Radboud Repository

Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/113607

Please be advised that this information was generated on 2017-12-06 and may be subject to change.

3324

MENS OF MACHINE?

Het lichaam-geest probleem in de cognitieve psychologie

M.A.M.M. Meijsing

MENS OF MACHINE?

het lichaam-geest probleem
in de cognitieve psychologie

MENS OF MACHINE?

het lichaam-geest probleem in de cognitieve psychologie

PROEFSCHRIFT

ter verkrijging van de graad van doctor
in de wijsbegeerte
aan de Katholieke Universiteit te Nijmegen,
op gezag van de Rector Magnificus Prof. Dr. J.H.G.I. Giesbers
volgens besluit van het College van Dekanen
in het openbaar te verdedigen
op donderdag 25 september 1986
des namiddags te 1.30 uur precies

door

Monica Antoinette Michaela Maria Meijsing

geboren te Haarlem

SWETS & ZEITLINGER B.V. / LISSE
1986

Promotores: Prof dr C.E.M. Struyker Boudier
Prof. dr. G.A.M. Kempen
Prof. dr. A.A. Derksen

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Meijsing, Monica Antoinette Michaela Maria

Mens of machine? : het lichaam-geest probleem in de cognitieve psychologie / Monica Antoinette Michaela Maria Meijsing. - Lisse : Swets & Zeitlinger
Proefschrift Nijmegen. - Met lit. opg.
ISBN 90-265-0760-7
SISO 415.4 UDC 159.95(043.3)
Trefw.: lichaam en geest ; cognitieve psychologie.

Omslag ontwerp H. Veltman
Gedrukt bij Offsetdrukkerij Kanters b.v., Alblasserdam
© Copyright 1986 Swets &Zeitlinger en M. Meysing

Alle rechten voorbehouden Niets uit deze uitgave mag worden verveelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand, of openbaar gemaakt, in enige vorm of op enige wijze, hetzij elektronisch, mechanisch, door fotokopieën, opnamen, of op enige andere manier, zonder voorafgaande schriftelijke toestemming van de uitgever

All rights reserved No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher

ISBN 90 265 0760 7

Voor Gerlof en voor mijn ouders.

INHOUDSOPGAVE

1. PROBLEEMSTELLING.

1.1. Inleiding.

Bij het bestuderen van de mens (en van sommige dieren) valt het onmiddellijk op dat zich twee verschillende soorten verschijnselen voordoen. Enerzijds is er de mens als menselijk lichaam, dat wil zeggen als iets dat beschouwd kan worden als een fysisch object met allerlei fysische eigenschappen, en anderzijds heeft diezelfde mens eigenschappen die geen enkel puur fysisch object lijkt te hebben. Een mens kan bewegen, vallen, iets breken, oplossen in ongebluste kalk enz; maar een mens kan ook iets bedenken of willen, pijn voelen, bang zijn voor de dood. Het onderscheid tussen wat wij het fysische en het mentale noemen dringt zich bij het nadenken over en het bestuderen van de mens als een in onze ervaring gegeven feit aan ons op. Men kan zich afvragen wat het fysische en wat het mentale is, wat beide onderscheidt, wat de relatie tussen beide is. Dit zijn filosofische vragen. Tesamen genomen vormen zij het terrein van het lichaam-geest probleem in ruimere zin. (De laatste van de drie genoemde vragen - die betreffende de *relatie* van lichaam en geest - wordt dan als het lichaam-geest probleem in engere zin aangeduid.) Ook de psychologie, als studie van (het gedrag van) de mens, kan dit probleem niet negeren.

De psychologie hoeft de filosofische vragen van het lichaam-geest probleem niet te thematiseren of er een oplossing voor te vinden - als dat al überhaupt mogelijk is - maar ze moet toch in haar uitgangpunten haar positie bepalen ten opzichte van het lichaam-geest probleem. Omgekeerd zijn de bevindingen van de psychologie als empirische wetenschap relevant voor de filosofische vragen en mogelijke oplossingen van het lichaam-geest probleem.

De psychologie kan zich aan deze positiebepaling ten opzichte van het lichaam-geest probleem onttrekken door te vermijden ooit naar het mentale te verwijzen en door haar hele domein in uitsluitend fysische termen te beschrijven. Het gedrag wordt dan gezien als een fysisch gevolg van fysische oorzaken, en zogenaamde mentalistische termen als 'denken', 'willen', 'gevoelens' worden vermeden. Een dergelijke

vermijding van mentalistische termen kan het gevolg zijn van een methodologische beslissing om de psychologie te beschouwen als een fysische wetenschap van gedrag, of ze kan het gevolg zijn van het aanhangen van een fysicalistische filosofie van het mentale, en wel van het eliminatief materialisme Dit eliminatief materialisme stelt dat mentalistische termen nergens naar verwijzen en daarom uit de taal geëlimineerd kunnen (en moeten) worden, het stelt dat het mentale niet bestaat, zoals bijvoorbeeld heksen of duivels niet bestaan

Een psychologie die het niet mag hebben over zulke alledaagse onderwerpen als meningen en wensen en gevoelens doet op zijn minst erg steriel en gekunsteld aan; niet ten onrechte wordt van de behavioristen die mentalistische termen meden wel gezegd dat ze anesthesie moesten voorwenden Mijn belangstelling gaat uit naar een niet-geanestheseerde psychologie, naar een psychologie die het wel over meningen en wensen en gevoelens heeft Het eliminatief materialisme, dat stelt dat zulke mentale verschijnselen, die toch zo'n belangrijke rol in ons leven lijken te spelen, niet bestaan, lijkt me dan ook een *counsel of despair.*

In deze studie wil ik de mogelijkheid onderzoeken van een psychologie die over meningen en wensen en gevoelens gaat; ik wil zoeken naar een filosofische positie ten aanzien van het lichaam-geest probleem (in ruimere zin) die onze alledaagse ervaring van zulke mentale verschijnselen, in eerste instantie althans, serieus neemt, en die zo'n psychologie kan funderen

De moderne cognitieve psychologie is zo'n psychologie die vrijelijk mentalistische termen gebruikt. De cognitieve psychologie beweert tevens een fysicalistische oplossing voor het lichaam-geest probleem te kunnen geven. De oplossing voor het lichaam-geest probleem die de cognitieve psychologie zegt te bieden is op het eerste gezicht zeer elegant en simpel. Men stelt "Computers, en meer nog robots, zijn net als mensen Ze kunnen intelligentie-vereisende taken verrichten; hun gedrag kan verklaard worden in termen van meningen en wensen Maar computers en robots zijn tevens machines, volledig fysische mechanismen. Ze zijn door mensen gemaakt, en hun gedrag kan verklaard worden in termen van fysische oorzaken De juist geprogrammeerde computer vormt een existentiebewijs voor de mogelijkheid dat twee soorten verklaringen, verklaringen in termen van

meningen en wensen èn verklaringen in termen van fysische oorzaken, van toepassing zijn op één volledig fysisch systeem. Daarom is het op zijn minst mogelijk dat mensen net als computers zijn: volledig fysische systemen waar twee soorten verklaringen op van toepassing zijn".

Deze fysicalistische oplossing van het lichaam-geest probleem is in haar eenvoud en elegantie zeer aantrekkelijk. Er wordt recht gedaan aan de fundamentele ervaring dat mensen meningen en wensen (en gevoelens) hebben en op grond daarvan handelen. Tegelijkertijd wordt de eenheid van de wetenschap gewaarborgd: alles is fysisch en verklaarbaar in termen van fysische oorzaken.

Ik wil deze oplossing onderzoeken, zowel zoals ze in filosofisch naïeve en goeddeels impliciete vorm leeft onder vele cognitieve psychologen, als in de expliciete uitwerking van een aantal filosofen. Mijn conclusie zal zijn dat deze fysicalistische theorie van het mentale niet houdbaar is en dat de cognitieve psychologie ook in feite niet deze theorie steunt of erdoor gefundeerd wordt. Een andere theorie van het mentale, de dubbelaspecttheorie, kan vele verworvenheden van die fysicalistische theorie van het mentale overnemen, en sluit beter aan bij de cognitieve psychologie.

Voor zover hier de mogelijkheid van een mentalistische psychologie in aansluiting op een consistente theorie van het mentale wordt aangetoond, vormt deze studie tevens een weerlegging van het eliminatief materialisme.

1.2. Opzet van het onderzoek.

De opzet van het onderzoek is als volgt:
In 1.3 vermeld ik kort de verschillende posities die mogelijk zijn ten aanzien van het lichaam-geest probleem. In 1.4 laat ik zien dat in de moderne cognitieve psychologie een fysicalistische positie wordt aangehangen. Toch is dit een mentalistische psychologie die intentionaliteit ziet als kenmerk van mentale.

Vervolgens toon ik aan hoe cognitieve psychologen een fysicalistische oplossing voor het lichaam-geest probleem zien. Ik doe dit aan de hand van de psychologe/filosofe Margaret Boden, die als een van de eersten in de cognitieve psychologie het belang van computers voor

verklaringen in de psychologie en voor het lichaam-geest probleem signaleerde. In de kritiek op haar oplossing worden twee dingen duidelijk: ten eerste steunt de oplossing op twee poten, een empirische poot en een *apriori* poot, en ten tweede behoeven zowel de empirische als de *apriori* poot een veel verdere uitwerking.

De empirische poot wordt gevormd door de uitspraak dat er nu computers bestaan die net als mensen zijn. In hoofdstuk 2 laat ik (na enige definities) zien dat deze uitspraak in twee opzichten problematisch is. Ten eerste moet nader gespecificeerd worden in welk opzicht computers net als mensen zijn, met ander woorden met welke *grain* of korrel men moet kijken om de gelijkheid (in cognitief opzicht) tussen computer en mensen te zien. In een historisch overzicht (2.3) bespreek ik dit *grain* probleem. Ten tweede zijn er redenen om aan te nemen dat het niet mogelijk is om globale cognitieve processen door de ons nu bekende computers te laten uitvoeren. Ik bespreek dit *frame* probleem in 2.4.

Mijn conclusie over de mens-machine gelijkheid is dat in ieder geval nu computers te weinig net als mensen zijn om zo'n zware rol te spelen in de empirische poot van de fysicalistische oplossing van het lichaam-geest probleem. Als men gelooft dat in de toekomst computers ook globale cognitieve processen kunnen uitvoeren, dan is dat omdat men op *apriori* gronden gelooft dat mensen volledig fysische systemen zijn, die derhalve nagebouwd moeten kunnen worden. Dit geloof in een fysicalistische oplossing van het lichaam-geest probleem kan zelf geen empirische steun vinden in het bestaan van computers, die immers niet voldoende op mensen lijken. De *apriori* poot moet nu de hele fysicalistische oplossing van het lichaam-geest probleem dragen.

Zoals gezegd wordt deze poot gevormd door een fysicalistische theorie van de mens en van het mentale. Met name twee punten in de naïeve versie van deze theorie behoeven nadere uitwerking ten eerste, de relatie tussen verklaringen van de mens in fysische termen en verklaringen in termen van meningen en wensen enz; en ten tweede, de vraag of bepaalde systemen zoals de mens (of de computer) echt meningen en wensen hebben of dat het enkel om heuristische redenen handig is om ze meningen en wensen instrumentalistisch toe te schrijven.

Deze punten zijn uitgewerkt in de nu gangbare filosofie van het

mentale: het functionalisme In hoofdstuk 3 schets ik de algemeen aanvaarde grondlijnen van dit functionalisme in contrast met het behaviorisme en de identiteitstheorie Daarin wordt de relatie tussen de bovengenoemde soorten verklaringen, en de relatie tussen de fysica en de speciale wetenschappen, uiteengezet. Het functionalisme gaat samen met het token-fysicalisme, dat stelt dat alle gebeurtenissen en objecten fysische gebeurtenissen en objecten zijn. Vervolgens worden twee *erkende problemen van het functionalisme besproken. het probleem* van het Turingmachine-functionalisme en het *qualia* probleem.

Na deze inleiding tot het functionalisme in hoofdstuk 3 worden de twee voornaamste varianten van het functionalisme onderzocht in de hoofdstukken 4 en 5. Deze varianten onderscheiden zich in hun antwoord op de vraag of sommige (fysische) systemen echt meningen en wensen hebben, de ene variant is realistisch en de andere instrumentalistisch op dit punt.

De realistische variant wordt in hoofdstuk 4 behandeld aan de hand van een analyse van het werk van Jerry Fodor. Ik heb hem uitgekozen omdat hij op dit moment de meest uitgewerkte en meest omvattende realistische theorie van het mentale geeft ter fundering van de cognitieve psychologie. Realistische posities die afwijken van de zijne komen in de bespreking van zijn argumenten aan de orde. Fodor's realisme ten aanzien van meningen en wensen (propositionele attitudes) en zijn erkenning van de intentionaliteit van meningen en wensen brengen hem tot het postuleren van interne representaties (4.2). Zijn fysicalistische theorie van mentale veroorzaking doet hem die representaties postuleren als expliciet aanwezige, fysische (neurale) entiteiten, die een causale rol spelen op grond van hun vorm.

Deze laatste stap levert hem drie samenhangende problemen op: wat maakt dat die neurale entiteiten representaties van de wereld zijn, wat maakt dat ze überhaupt iets (eventueel niet bestaands) representeren, en voor wie representeren ze iets (4.5)? Fodor onderscheidt deze problemen te weinig, en trekt ze samen tot het probleem van de semantische interpretatie van de interne representaties. Ik bespreek (in 4.6) twee pogingen om dit probleem op te lossen die Fodor zelf afwijst, de procedurele semantiek en de functionele rol semantiek, en twee pogingen tot een oplossing van hemzelf. Mijn conclusie is dat geen van die semantische theorieën de drie problemen oplost. Geen

enkele poging kan een unieke interpretatie van de representaties vastleggen in puur fysicalistische termen, en zelfs als dat wel kon, dan blijft het probleem dat er een instantie moet zijn voor wie de representaties die interpretatie hebben Zolang dit probleem niet is opgelost blijft intentionaliteit als kenmerk van het mentale vooronderstelt en niet verklaard, zolang is er dan geen fysicalistische theorie van het mentale.

In hoofdstuk 5 bespreek ik de instrumentalistische variant van het functionalisme aan de hand van Daniel Dennett Ik heb hem uitgekozen omdat hij een uitgewerkte en veelomvattende theorie van het mentale geeft en de belangrijkste representant is van het instrumentalisme met betrekking tot meningen en wensen. Ofschoon hij van mening is dat meningen en wensen enz uiteindelijk niet bestaan, verschilt zijn positie in zoverre van het eliminatief materialisme dat hij stelt dat het *toeschrijven* van meningen en wensen predictieve successen oplevert, en in de psychologie dus toegestaan is.

Eerst schets ik Dennett's instrumentalistische theorie en zijn redenen voor zo'n instrumentalisme (5.2 en 5 3) Hij ziet het probleem dat interne representaties een instantie vooronderstellen voor wie ze iets representeren. Vervolgens probeer ik de vraag te verhelderen of Dennett inderdaad, zoals vaak gedacht wordt, intentionaliteit wil verklaren in termen van complexiteit, net als intelligentie Dennett is hierover niet erg duidelijk. Maar men kan hem ook anders lezen In een doorgevoerde instrumentalistische lezing van zijn theorie hoeft hij intentionaliteit niet in termen van complexiteit te verklaren, en doet hij dat ook niet (5 4) Tenslotte illustreert een uitgebreid voorbeeld zijn theorie (5.5).

Daarna bekritiseer ik Dennett's theorie De argumenten die hij geeft om bepaalde mentale entiteiten te elimineren zijn verre van steekhoudend, en de extreme derde-persoons positie die hij moet innemen om zijn theorie consistent te kunnen verdedigen is ronduit verificationistisch Mijn voornaamste kritiekpunt is evenwel dat in Dennett's eigen voorbeeld van een fysisch model van de mens nog steeds intentionaliteit wordt vooronderstelt, en dat zelfs in de extreme derde-persoons lezing van zijn instrumentalistische theorie de intentionaliteit van degene die meningen en wensen aan anderen (en zichzelf!) *toeschrijft* onverklaard blijft Omdat zijn theorie altijd nog

ergens intentionaliteit vooronderstelt, is het geen fysicalistische theorie van het mentale

In hoofdstuk 6 stel ik een andere theorie van het mentale voor: de dubbelaspecttheorie Ik laat eerst zien dat het vooronderstellen van intentionaliteit in plaats van het te verklaren alleen problematisch is als men een fysicalistische theorie wil (6.1). Empirische aanwijzingen voor het token-fysicalisme zijn er niet, noch directe (6.2) noch indirecte (6.3). De cognitiewetenschap steunt in feite niet het token-fysicalisme, en vooronderstelt het evenmin.

Wat Fodor en Dennett voorstellen is een aanvulling op of een hervorming van de cognitieve psychologie van een intentionele tot een fysicalistische theorie Maar als zij er niet in slagen zelf een consistente fysicalistische theorie van het mentale te formuleren, lijkt het me beter om een andere theorie van het mentale te accepteren die wel consistent is en die wel bij de huidige cognitieve psychologie aansluit (6.4). De dubbelaspecttheorie kan vele verworvenheden van het token-fysicalisme overnemen en heeft minder problemen (6.5). Wat aan de cognitieve psychologie hervormd moet worden als men de dubbelaspecttheorie accepteert is niet het werk zelf, maar alleen de opvatting over wat dat werk is, het zelfbegrip van de psychologie.

1.3. Schets van het lichaam-geest probleem.

Het lichaam-geest probleem in ruime zin bestaat, zoals gezegd, uit drie vragen: wat is het fysische en wat is het mentale, wat onderscheidt beide en wat is de relatie tussen beide. Voordat men de laatste vraag (het lichaam-geest probleem in engere zin) kan beantwoorden, moeten eerst de beide andere vragen beantwoord zijn. In het dagelijks leven spreekt men van gegevenheden in het menselijk lichaam, zoals de schedel, bloedvaten, spijsvertering, die fysisch (materieel) genoemd worden, en van andere gegevenheden, zoals hoofdpijn en een gedachte over aspirine, die men psychisch (mentaal) noemt. Er is dus sprake van twee verschillende soorten gegevenheden.

Men kan zich afvragen: wat wordt precies bedoeld met 'twee verschillende soorten gegevenheden'? Waarom *twee verschillende* soorten gegevenheden? Dat zou betekenen dat alle gegevenheden van

één soort iets met elkaar gemeen hebben dat ze niet gemeen hebben met de gegevenheden van de andere soort. Is dat zo? Niet helemaal Men kan, zo neutraal mogelijk, stellen dat de fysische gegevenheden alle beschreven kunnen worden in termen van de fysica. Maar kunnen alle psychische gegevenheden beschreven worden in termen van een theorie, één psychologie of één theorie van het mentale? Dat is (nog) niet het geval. De psychische gegevenheden worden zelf vaak weer onderverdeeld in twee soorten. enerzijds de sensaties (van pijn, van kleur, van stemming enz), anderzijds de zogenaamde psychologische (of propositionele) attitudes, zoals meningen, wensen, verwachtingen enz., die gekenmerkt worden door intentionaliteit (zie 1.4.2) Theorieën van het mentale gaan zelden over beide soorten mentale gegevenheden (zie verder ook 3.4.2). Wanneer we even afzien van dit onderscheid binnen de mentale gegevenheden, moeten we ons nog afvragen wat de psychische (of mentale) gegevenheden onderscheidt van de fysische. Op deze vraag zijn drie verschillende antwoorden mogelijk.

Het *eerste* antwoord is. niets onderscheidt de mentale gegevenheden van de fysische, want één van beide soorten gegevenheden bestaat niet. Van dit antwoord bestaan weer twee varianten Enerzijds is er het *eliminatief materialisme* dat stelt dat mentale gegevenheden niet bestaan. Onze taal kent wel mentale termen, maar die verwijzen nergens naar, zoals ook de term 'phlogiston' nergens naar verwijst. In de toekomst zullen we alles wat we zouden willen zeggen in een puur fysicalistische taal kunnen zeggen, want het fysische is het enige dat bestaat (b.v. Feyerabend 1970, Rorty 1971b, Churchland 1979, 1984). Anderzijds is er de extreem *idealistische positie* dat fysische gegevenheden niet bestaan: alles wat er is zijn mentale gegevenheden, en fysische gegevenheden, met name objecten, zijn slechts geconstrueerd uit onze sensaties (b.v. de positie van Berkeley). Beide posities menen dat één soort gegevenheden op illusies berust. Voor de eliminatieve materialisten is dat problematisch omdat een illusie zelf een mentaal verschijnsel is. De extreem idealistische positie leidt tot een volledig solipsisme, tenzij men gelooft in de goedheid van (een) God die een wereld laat corresponderen met onze sensaties.

Het *tweede* antwoord op de vraag wat de psychische gegevenheden onderscheidt van de fysische luidt: niets wezenlijks. Men meent dan

dat het onderscheid louter behoort tot de naieve ervaring, en dat in werkelijkheid, 'achter die ervaring, een identiteit van het psychische met het fysische bestaat Deze positie is in verschillende vormen van het *materialisme* uitgewerkt Het *metafysisch behaviorisme* stelt, in tegenstelling tot het eliminatief materialisme, dat mentale termen wel ergens naar verwijzen Ze verwijzen echter niet naar mentale entiteiten of processen of toestanden, maar naar (categorieen van) (fysisch) gedrag Dit behaviorisme stelt dat zinnen in mentale termen hetzelfde betekenen als, en dus vertaald kunnen worden in, zinnen in termen van gedrag(sdisposities) (b v Ryle 1949, Skinner 1974) (1) Een probleem voor deze positie is dat iemand gedrag kan vertonen zonder dat de bijbehorende mentale term van toepassing is, bijvoorbeeld in het geval van een acteur die woede moet voorwenden (zie ook Austin s vrolijke artikel Pretending (1967) waarin hij het standpunt ridiculiseert dat er bij woede-gedrag, als het maar ver genoeg gaat, echt sprake van woede moet zijn) De *identiteitstheorie* stelt dat mentale toestanden en gebeurtenissen identiek zijn aan neurofysiologische toestanden en gebeurtenissen Dat wil niet zeggen dat zinnen in mentale termen hetzelfde betekenen als zinnen in neurofysiologische termen, ze verwijzen alleen naar dezelfde gebeurtenissen en toestanden De geclaimde identiteit is een empirisch ontdekte, contingente identiteit (b v Place 1956, Smart 1959, Armstrong 1968) Deze positie is problematisch omdat niet duidelijk is wat identiek is aan wat Wanneer twee dingen identiek zijn dan is er toch maar een ding? Wanneer we zeggen dat Shakespeare identiek is aan Bacon, dan moeten we toch zeggen dat er helemaal afzonderlijk individu Shakespeare bestaat (of geen Bacon!), dus hoe kan dan zo iemand identiek zijn aan Bacon? Volgens dit argument moet de identiteitstheorie overgaan in een eliminatief materialisme degene die mensen abusievelijk Shakespeare noemden bestaat niet als afzonderlijk individu, alleen Bacon bestaat en heeft een aantal (maar niet alle) eigenschappen van de vermeende Shakespeare, datgene wat mensen abusievelijk mentale toestanden' noemden bestaat niet, alleen fysische toestanden bestaan en hebben een aantal (maar niet alle) eigenschappen van de vermeende mentale toestanden

Wanneer men deze stap naar het eliminatief materialisme niet wil maken blijft men zitten met een dualisme van talen of van wijzen van

kennen (b v Feigl 1958) mentale toestanden zijn hersentoestanden, maar we spreken er in twee talen over, een mentalistische en een fysicalistische, en we hebben er twee manieren van kennen voor, 'van binnen uit' voor degene die de toestand heeft, en 'van buiten af' voor de hersenonderzoeker

Zo'n dualisme van wijzen van kennen leidt tot een volgende vraag: als we twee wijzen van kennen hebben van dezelfde toestand moeten we dan ook niet spreken van twee heel verschillende soorten van eigenschappen *waaraan* we die toestand op twee heel verschillende manieren kennen? Is dat niet een dualisme van eigenschappen? Ik ken mijn nabeeld als bijvoorbeeld rood, een hersenonderzoeker kent diezelfde toestand (mijn nabeeld is immers een hersentoestand volgens de theorie) als bijvoorbeeld hoogfrequent vurend. Volgens de wet van Leibniz moeten twee dingen die identiek blijken te zijn ook al hun eigenschappen gemeenschappelijk hebben. Maar kun je wel spreken van een rode hersentoestand (mijn hersenen zijn grijs) of van een hoogfrequent vurend nabeeld (mijn nabeeld is egaal rood)? (Zie voor discussies over dit probleem b.v Borst 1970, Sellars 1971, Rosenthal 1971.)

Tegen de claim dat de identiteit tussen mentale toestanden en hersentoestanden een *contingente* identiteit is argumenteerde Kripke (1972) dat alle identiteitsclaims gesteld zijn in termen van zogenaamde *rigid designators*. Deze duiden dezelfde entiteit aan in alle mogelijke werelden. Derhalve kunnen de identiteitsclaims alleen *noodzakelijk* waar zijn, maar dat zijn ze, volgens de verdedigers van de claim, niet. Kripke concludeert dat de identiteitsclaims dan helemaal niet waar kunnen zijn. Een ander probleem voor de identiteitstheorie is empirisch; dezelfde mentale processen en toestanden gaan vaak samen met verschillende hersenprocessen en -toestanden Bijvoorbeeld bij lesies in de hersenen blijkt vaak dat een ander deel van de hersenen de 'taak' van het beschadigde deel overneemt; bepaalde mentale toestanden zijn dan in een ander deel van de hersenen gelocaliseerd. Bovendien kan men niet uitsluiten dat ook wezens met radicaal anderssoortige hersenen (marsmannetjes?) mentale toestanden hebben. De identiteitstheorie is mede vanwege dit probleem door sommigen verfijnd tot of vervangen door het *functionalisme*. (Het functionalisme wordt uitvoerig besproken in hoofdstuk 3, 4 en 5.)

Het *derde* antwoord op de vraag wat de psychische gegevenheden onderscheidt van de fysische luidt: iets wezenlijks (b.v. intentionaliteit, zie 1.4.2). Men meent dan dat het verschil onherleidbaar is. Er is dan in ieder geval sprake van een dualisme van gegevenheden. De volgende vraag die nu gesteld moet worden is: wat voor soort dualisme wordt aangehangen?

Men kan kiezen voor een *dualisme van toestanden of van eigenschappen*. Dan moet de vraag beantwoord worden: wie of wat heeft die twee verschillende soorten toestanden of eigenschappen? Daar zijn weer verschillende antwoorden op mogelijk. Men kan zeggen: een volledig fysisch systeem kan twee soorten eigenschappen of toestanden hebben. Deze positie neigt ertoe over te gaan tot een fysicalisme, wanneer men bedenkt dat alle toestanden van een volledig fysisch systeem toch fysische toestanden zijn (zie ook de bespreking van de identiteitstheorie en van Leibniz' wet hierboven). Wanneer men evenwel ontkent dat de mentale toestanden en eigenschappen fysische toestanden en eigenschappen zijn, neigt de positie tot een *epifenomenalisme:* aangezien het fysisch systeem volgens volledig fysische wetten werkt zijn de mentale toestanden en eigenschappen er als een bijverschijnsel, een epifomeen, opgeplakt, en maakt hun aanwezigheid geen verschil in de wereld.

Op de vraag wie of wat die twee soorten toestanden of eigenschappen heeft kan men ook antwoorden: de persoon. De persoon heeft dan zowel fysische als psychische eigenschappen. Men moet dan de notie van 'persoon' als primitief beschouwen, en niet meer willen vragen of de persoon fysisch of psychisch is of een combinatie van beide (b.v. Strawson 1959). Deze positie, een dubbelaspecttheorie of persoonstheorie genoemd, zet zich dus af tegen zowel de (fysicalistische) identiteitstheorie àls tegen het Cartesiaans dualisme (zie hieronder). Ze doet recht aan de fundamentele intuitie van de eenheid van de persoon (Om een voorbeeld te geven: Mijn gewicht en mijn zorgen daarover zijn beide eigenschappen van mij). De dubbelaspecttheorie verzet zich ook tegen de sinds Descartes gangbare opvatting dat het menselijk lichaam een materieel object, een machine, zou zijn, gestuurd door een geest. Problemen met het Cartesiaans dualisme (zie sectie II) gaven vervolgens mede aanleiding tot de opvatting dat de mens helemaal een materieel object zou zijn, een

machine zonder geest.

Mijn lichaam is niet een ding dat ik heb en gebruik, of een automaat, een machine; ik ben het zelf, ik doorleef mijn hele lijf (zie voor dit punt b.v. continentale filosofen als Husserl, Scheler, Merleau-Ponty; zie ook b.v. Dekkers 1985) (2). Ook het lichaam van een ander kan men hooguit in een soort neutrale laboratoriuminstelling opvatten als een materieel object. De lichamen van vrienden en minnaars, van vijanden, maar ook van volslagen vreemden in een overvolle bus, zijn allemaal eigen lijven (3); slechts in een abstractie kan men ze beschouwen als fysische objecten. Een persoon is niet identiek met zijn lichaam als fysisch object (4), en men kan ook niet zeggen dat de persoon zijn lichaam (als fysisch object) heeft, bezit. Een persoon is een entiteit die noodzakelijk zowel psychische als fysische eigenschappen heeft; zowel een psychisch als een fysisch aspect; de persoon zelf is noch fysisch noch psychisch. Je kunt hem vanuit een bepaalde, fysicalistische theorie, beschouwen als een fysisch object, een machine, maar je moet je dan wel bewust zijn dat dat een abstractie is. Je kunt hem ook, vanuit een psychologische theorie, beschouwen als een psychisch wezen, maar ook dat is een abstactie, niet het hele plaatje. Beide soorten theorieën zijn manieren van kijken.

De dubbelaspecttheorie doet ook recht aan de fundamentele intuïtie van de meeste mensen dat er een categoriaal verschil is tussen personen en dingen. Dingen - tafels, stoelen, stenen - hebben maar één soort eigenschappen: fysische eigenschappen. Dingen zijn wel fysische objecten, het zijn in ieder geval geen personen. Waar precies in de fylogenetische schaal tussen planten en mensen de grens tussen dingen en personen gelegd moet worden is natuurlijk problematisch. Voor een panpsychist kan de hele wereld uit personen bestaan; zo'n positie is ook verenigbaar met een dubbelaspecttheorie. Maar voor de meeste mensen zijn stenen en thermostaten geen personen, en mensen wel, en honden en katten en paarden misschien een beetje (zie voor de opvatting dat de grens tussen dingen en personen steeds meer naar de mens toe 'omhoog' is geschoven Sellars 1971).

Tenslotte kan men op de vraag wie of wat die twee soorten toestanden of eigenschappen heeft antwoorden: twee soorten entiteiten. Het lichaam heeft fysische eigenschappen en de geest heeft psychische eigenschappen. Dit is de sterkste vorm van dualisme, het *Cartesiaans*

dualisme. Volgens deze positie bestaat de wereld uit twee soorten entiteiten materie(le objecten) en geest De mens is een compositum van beide Een probleem voor deze positie is duidelijk te maken wat een geestelijke substantie of entiteit die geen enkele fysische eigenschap heeft kan zijn Een ander probleem is dat de eenheid van de mens verloren gaat de mens is een (ondenkbaar) samenstelsel van twee entiteiten

Wanneer men gekozen heeft voor een bepaald antwoord op de vraag wat bedoeld wordt met lichaam en geest, en met twee verschillende soorten gegevenheden, komt men toe aan de vraag wat is de relatie tussen lichaam en geest? Deze vraag wordt wel het eigenlijke lichaam-geest probleem of het lichaam-geest probleem in engere zin genoemd De vraag wat de *relatie* tussen lichaam en geest is geldt eigenlijk alleen voor het Cartesiaans dualisme, dat immers als enige positie stelt dat lichaam en geest onderscheiden en scheidbare enititeiten zijn De enorme moeilijkheid van deze vraag is voor velen een motief geweest om een andere positie te kiezen Voor alle monistische posities doet de vraag zich niet voor, evenmin als voor de persoonstheorie De Cartesiaanse dualist kan op de vraag wat is de relatie tussen lichaam en geest drie antwoorden geven

Het meest voor de hand ligt het *eerste* antwoord, *interactionisme.* Deze positie van wederzijdse beinvloeding sluit direct aan bij ons alledaags taalgebruik ik voelde pijn omdat ik in een spijker trapte (fysische gebeurtenis veroorzaakt psychische gebeurtenis), en ik riep "au" omdat ik pijn had (psychische gebeurtenis veroorzaakt fysische gebeurtenis) (b v Popper en Eccles 1977) Maar kan dat eigenlijk wel? Als het fysische en het psychische, lichaam en geest, zo totaal verschillend zijn, hoe kunnen ze dan elkaar beinvloeden? Als het psychische een fysische verandering veroorzaakt, moet het dan niet zelf een fysische oorzaak, dus fysisch zijn? Bovendien, meer en meer wordt in de natuurwetenschappen de geslotenheid van de fysische wereld benadrukt De wet van behoud van energie stelt dat er in de fysische wereld nooit energie bijkomt of verloren gaat Dat maakt interacie met een niet-fysische entiteit onmogelijk Interactionisten hebben uitwegen gezocht voor deze onmogelijkheid door te stellen dat de wet van behoud van energie misschien minimale afwijkingen toelaat die op langere duur uitmiddelen Een andere strategie is om de geest

te laten aangrijpen op het niveau van de quantummechanica, waar de locatie van de subatomaire deeltjes niet gedetermineerd is (b v Margenau 1977).

Vanwege de problemen met de wet van behoud van energie in de fysische wereld kan men ook een *tweede* antwoord geven en zeggen dat het lichaam wel de geest beïnvloedt, maar de geest niet het lichaam. Voor een verandering van psychische toestand is geen fysische energie nodig, dus er gaat geen energie verloren; en de geest werkt niet op het lichaam dus er komt geen energie bij. Dit is een variant van het bovengenoemde *epifenomenalisme.* Het enige verschil is dat hier nog een geestelijke substantie wordt aangenomen als drager van de psychische toestanden en eigenschappen. Het probleem voor het epifenomenalisme is ook hier dat de geest (en al het psychische) er niets toe doet, zonder dat zou de wereld precies hetzelfde verlopen.

Een *derde* antwoord op de vraag wat de relatie is tussen lichaam en geest luidt. er is geen beïnvloeding. Psychische gebeurtenissen veroorzaken alleen psychische gebeurtenissen en fysische gebeurtenissen veroorzaken alleen fysische gebeurtenissen. Volgens dit *parallellisme* lopen beide soorten gebeurtenissen volstrekt parallel, zoals twee klokken parallel lopen, dezelfde tijd aangeven, zonder elkaar te beïnvloeden. Het probleem voor deze positie is dat er geen gesloten causale keten van louter psychische gebeurtenissen lijkt te bestaan Welke *psychische* gebeurtenis is er de oorzaak van dat ik pijn voel als ik in een spijker trap? De parallellie tussen het fysische en het psychische zonder dat er sprake is van enige interactie is wonderbaarlijk en onverklaarbaar. En de verklaring dat God de parallellie van te voren zo geregeld heeft (de vooraf vastgestelde harmonie van Leibniz), of zelfs bij iedere voorkomende gelegenheid zelf verzorgt (het occasionalisme van Malebranche) is voor weinigen acceptabel.

Het bovengegeven overzicht van het lichaam-geest probleem is zeer summier. De verschillende posities zijn in hun eenvoudigste vorm beschreven. Het ging er mij slechts om een kort overzicht te schetsen van de verschillende vragen en mogelijke antwoorden met betrekking tot het lichaam-geest probleem. In deze studie wordt de fysicalistische (materialistische) positie onder de loep genomen, en dan wel met name het functionalisme In de moderne psychologie wordt het fysicalisme als

vanzelfsprekend beschouwd. Zoals Dennett opmerkt.

> ". . it is widely granted these days that dualism is not a serious view to contend with, but rather a cliff over which to push one's opponents . ." (Dennett 1978a, 252).

1 4 Fysicalisme en psychologie.

In de moderne, wetenschappelijke psychologie beschouwt men het fysicalisme als vanzelfsprekend, zoals gezegd. Dat lijkt problematisch omdat de psychologie toch over het psychische of het mentale gaat. De nieuwste stroming in de wetenschappelijke psychologie, de cognitieve psychologie, gebruikt vele mentale termen en beschouwt zichzelf niet meer als de studie van enkel het gedrag Ze meent evenwel dat de studie van het psychische toch fysicalistisch kan zijn, en claimt een fysicalistische oplossing te kunnen bieden voor het lichaam-geest probleem

In de volgende paragrafen zal ik die oplossing, zoals die veelal impliciet en soms expliciet in de cognitieve psychologie wordt aangehangen, uiteenzetten. In 1 4.1 zal ik kort aangeven wat cognitieve psychologie is, in 1.4 2 bespreek ik wat men beschouwt als het kenmerk van het psychische dat in de cognitieve psychologie bestudeerd wordt, namelijk intentionaliteit, en in 1.4.3 laat ik zien hoe men meent dat het bestaan van het psychische, en van intentionaliteit als het kenmerk daarvan, geen afbreuk doet aan de basisthese van het fysicalisme.

1 4.1 Wat is cognitieve psychologie?

Een eenduidig antwoord op de vraag "Wat is cognitieve psychologie?" valt niet gemakkelijk te geven Cognitieve psychologie is een wetenschapsgebied dat gekenmerkt wordt door allerlei aannamen, uitgangspunten, methoden en discussies die in het volgende uiteengezet zullen worden Wanneer we de cognitieve psychologie, of

liever de hele cognitiewetenschap (zie onder) zien als een Kuhniaans paradigma, zoals wel gedaan is (b v. Hayes 1978), dan kunnen we zeggen dat het paradigma door al die aannamen, uitgangspunten methoden en discussies nooit volledig gedefinieerd kan worden (Kuhn 1962). En over wat wel expliciet geformuleerd is bestaat geen volledige consensus.

Toch is het wel mogelijk om iets te verduidelijken van wat cognitieve psychologie is. Cognitieve psychologie is een onderdeel van de cognitiewetenschap Cognitiewetenschap houdt zich bezig met de studie van cognitieve (kennisdragende, kennisverwerkende) systemen (zie b.v. Kempen 1983, 1984).

Een paar citaten kunnen helpen de eerste kennismaking met de cognitiewetenschap wat verder te brengen

> "Cognitive science, it is said, is a discipline emerging from the intersection of psychology, linguistics, artificial intelligence and the philosophy of mind" (Stich 1982a, 419).

> "Cognitivism in psychology and philosophy is roughly the position that intelligent behavior can (only) be explained by appeal to internal "cognitive processes", that is rational thought in a very broad sense" (Haugeland 1978a, 215).

De cognitieve psychologie kan gezien worden als een reactie op het behaviorisme. Het behaviorisme stond, met zijn voornamelijk zeer eenvoudige leerexperimenten (eenvoudig in de zin dat de leeropdrachten eenvoudig waren, niet de opzet van de experimenten), waarbij de responsmogelijkheden drastisch beperkt waren, wel erg ver af van de alledaagse *folk psychology,* die gedrag in het dagelijks leven verklaart in termen van redenen, van wensen en meningen en doelstellingen. De cognitiewetenschap had sterk behoefte aan een mentalistisch vocabulaire omdat menselijk gedrag zo duidelijk vaak *niet* stimulus-gebonden is. Het extreme environmentalisme van het behaviorisme met zijn S-R paradigma voldeed eenvoudigweg niet (zie b.v. Koch 1964).

In 1960 verscheen het invloedrijke boek *Plans and the Structure of Behavior* van Miller, Gallanter en Pribram, dat een duidelijke breuk

betekende met het S-R paradigma Men ziet 1960 dan ook wel eens als het begin van de cognitiewetenschap Ofschoon de geboorte van de cognitiewetenschap vaak rond 1960 gedateerd wordt, is de discussie over de cognitiewetenschap, over de grondslagen en de identiteit van de cognitiewetenschap, pas twintig jaar later, rond 1980, in volle gang gekomen Eerdere pogingen om het eigen vakgebied zo expliciet te karakteriseren zijn niet of nauwelijks te vinden Misschien is dit het geval door wat Johnson-Laird in 1980 als een mogelijkheid beschrijft

> "One attitude - an optimistic one - is that cognitive science already exists and is alive and flourishing in academe we have all in our different ways been doing it for years The gentleman in Moliere s play rejoiced to discover that he had been speaking prose for forty years without realizing it perhaps we are merely celebrating a similar discovery" (Johnson-Laird 1980, 71)

Enige uitzonderingen daargelaten, komt de filosofische discussie (twee van de drie bovengenoemde citaten zijn van filosofen afkomstig) twintig jaar te laat op gang, en sommige psychologen reageren daar geergerd op, bijvoorbeeld

> 'As one of your thoroughly modern mentalists, I find it nice to be told that what I am doing is all right, certified correct by a qualified, licensed philosopher" (Norman 1980b, 90)

1 4 2 Intentionaliteit

In 1 2 zagen we dat men zich kan afvragen wat het bindend kenmerk van het psychische of het mentale is, dat het onderscheidt van het fysische De filosoof/psycholoog Franz Brentano meende daar in de vorige eeuw een antwoord op te kunnen geven Volgens Brentano is intentionaliteit het kenmerk dat het psychische onderscheidt van het fysische Intentionaliteit is bij hem een eigenschap van bewustzijnsacten, en wel de eigenschap van het gericht-zijn op een object Zo ben ik mij altijd *van iets* bewust, kan ik niet liefhebben

zonder *iemand of iets* lief te hebben, kan ik niet verlangen zonder *iets* te verlangen, niet hopen zonder *iets* te hopen Nu is het zo dat volgens Brentano datgene of diegene waarvan of van wie men zich bewust is, die men liefheeft enz., niet in werkelijkheid hoeft te bestaan Ik kan verlangen naar een pijnstiller die onmiddellijk helpt tegen hoofdpijn zonder enige bijwerking, maar zo'n pijnstiller hoeft niet te bestaan. Het object van een bewustzijnsact is bij Brentano dan ook niet een werkelijk bestaand object in de buitenwereld, onafhankelijk van de bewuste persoon. Nee, de bewustzijnsact is gericht op een intentioneel object. Het intentionele object is immanent, *in* het bewustzijn, binnen, en niet in de buitenwereld. Brentano neemt wel een buitenwereld, een werkelijkheid, aan. Zo'n werkelijkheid moet er doorgaans wel zijn als oorzaak voor het intentionele object, al hoeft er dus niet altijd iets in die werkelijkheid te corresponderen met het intentionele object. Het gaat hier evenwel om een verborgen, een gesupponeerde werkelijkheid. Weliswaar is iedere act gericht op een object, maar dat is het intentionele object; het werkelijke object buiten wordt niet bereikt, niet gekend In schema:

act ⟶ intentioneel object (⟵ werkelijk object)

Brentano's intentionaliteitsbegrip is in de continentale filosofie bekritiseerd en verder ontwikkeld. Een centraal probleem in zijn theorie is de kwetsbaarheid ervan voor een skeptische vraag. "Hoe weet je of er uberhaupt een buitenwereld is die het intentionele object veroorzaakt?" En het is ook problematisch hoe die relatie tussen werkelijk object en intentioneel object causaal kan zijn (5) Op de ontwikkelingen in de continentale filosofie wil ik hier niet verder ingaan. Het is Brentano's notie van intentionaliteit, en niet die van latere continentale filosofen, die in de cognitiewetenschap een rol is gaan spelen.

Ook in de cognitiewetenschap neemt men aan dat intentionaliteit het kenmerk van het psychische is. Alleen ziet men intentionaliteit niet meer als kenmerk van àl het psychische. Niemand kan ontkennen dat sommige processen en toestanden die we psychisch of mentaal noemen helemaal niet intentioneel zijn Een plotselinge pijn, een onbestemd gevoel van onbehagen, een naamloze angst, zijn nergens op gericht,

zijn geen intentionele toestanden

Het is gebruikelijk om een onderscheid te maken tussen mentale toestanden en processen die gekenmerkt zijn door intentionaliteit, zoals meningen en wensen, en die welke gekenmerkt worden door een onmiddellijk ervaren kwaliteit, zoals sensaties van pijn of kleur. In het laatste geval spreekt men wel van *qualia* of van *raw feels* (zie b v. Dennett 1969, Fodor 1981b, McGinn 1982a) (6). De cognitiewetenschap concentreert zich vooral op die mentale toestanden en processen die gekenmerkt worden door intentionaliteit, op meningen, wensen, vewachtingen enz.

Brentano's notie van intentionaliteit is ingevoerd in de Angelsaksische filosofie door Chisholm (1957, 1967) Chisholm haalt Brentano aan waar deze stelt dat psychologische fenomenen gekenmerkt worden door de gerichtheid op een object, en de 'intentionele inexistentie' van dat object De fenomenen die dit concept van intentionele inexistentie het duidelijkst illustreren zijn de psychologische (of propositionele) attitudes, bijvoorbeeld verlangen, hopen, wensen, zoeken, menen en aannemen Wanneer Brentano zegt dat deze attitudes intentioneel een object in zichzelf bevatten, verwijst hij naar het feit dat ze altijd een object hebben, ook al bestaan die objecten niet in de werkelijkheid (Chisholm 1957, 169). Fysische (niet-psychologische) fenomenen kunnen niet een object intentioneel in zichzelf bevatten. En dan zegt Chisholm:

> "These points can be put somewhat more precisely by referring to the language we have used ... We can formulate a working criterium by means of which we can distinguish sentences that are intentional, or are used intentionally, in a certain language from sentences that are not" (Chisholm 1957, 170)

Chisholm laat dan zien dat intentionele zinnen bepaalde logische kenmerken hebben. Iedere intentionele zin heeft niet altijd al die kenmerken, maar wel minstens een

In technische termen zijn deze kenmerken de volgende. referentiële opaakheid, het falen van existentiele generalisatie en de onmogelijkheid van implicatie van enige ondergeschikte bijzin (of de negatie ervan).

Een aantal voorbeelden het object van een wens kan niet zomaar aangeduid worden met verschillende termen die dezelfde referentie hebben Oedipus wilde met Iokaste trouwen, maar hij wilde beslist niet met zijn moeder trouwen Toch verwijzen 'Iokaste' en 'Oedipus moeder naar dezelfde vrouw (referentiele opaakheid)

Het object van een wens kan aangeduid worden met een term die helemaal nergens naar verwijst 'Hij wil een eenhoorn" impliceert niet dat er ook maar een eenhoorn bestaat, terwijl een niet-intensionele zin als "Hij berijdt een eenhoorn" alleen waar kan zijn als er echt een eenhoorn bestaat (geen existentiele generalisatie)

En tenslotte, "Hij wil dat de koningin op zijn promotie komt impliceert noch dat ze komt, noch dat ze wegblijft (geen implicatie van de ondergeschikte bijzin of van de negatie daarvan) Chisholm zegt dan

> "We may now re-express Brentano s thesis - or a thesis resembling that of Brentano - by reference to intentional sentences Let us say (1) that we do not need to use intentional sentences when we describe nonpsychological phenomena, we can express all of our beliefs about what is merely "physical" in sentences which are not intentional But (2) when we wish to describe perceiving, assuming, believing, knowing, wanting, hoping, and other such attitudes, then either (a) we must use sentences which are intentional or (b) we must use terms we do not need to use when we describe nonpsychological phenomena' (Chisholm 1957, 172-173)

Naar aanleiding van deze kenmerken van intentionele zinnen, wordt intentionaliteit ook wel eens gezien als eigenschap van zinnen of linguistische contexten In dat geval wordt de term ook wel gespeld met een 's' intensionaliteit'. Nu zijn er zinnen waarvan de betekenis niets te maken heeft met de gerichtheid op een object, en die dus niets met Brentano's notie van intentionaliteit te maken hebben, maar die wel alle logische kenmerken van intentionele zinnen hebben Voorbeelden zijn modale zinnen die niets psychologisch hebben "Het is noodzakelijk dat als Polonius de man achter het scherm was, Polonius

de man achter het scherm was" of "Het is mogelijk dat longkanker veroorzaakt wordt door roken". Deze voorbeelden voldoen respectievelijk aan de criteria van referentiele opaakheid, geen implicatie van de ingebedde bijzin en falen van existentiele generalisatie Maar het is duidelijk dat zinnen die gaan over het gericht zijn van een persoon op een object, en die dus intentioneel zijn in de psychologische zin, ook intentioneel zijn in de logische zin.

De verschillende betekenissen en spellingen van het woord 'intentionaliteit' in het Angelsaksische taalgebruik zijn niet altijd even duidelijk. Soms wordt het met een 's' gespeld om het te onderscheiden van het begrip 'intentionality' dat 'opzettelijkheid' betekent, en te maken heeft met 'intention', 'bedoeling'. Anderen schrijven het om die reden met een hoofdletter.

Sommigen spellen het woord met een 's' om aan te geven dat ze een eigenschap van zinnen bedoelen, en wel non-extensionaliteit. Extensionaliteit is een begrip uit de logica; *extensioneel* noemt men een samengestelde zin waarvan de waarheidswaarde alleen afhangt van de waarheidswaarde van de samenstellende delen. Veelal wordt er geen onderscheid gemaakt tussen intentionaliteit als eigenschap van zinnen en als eigenschap van het psychologische, of de geest.

De taalfilosoof John Searle heeft zich onlangs sterk verzet tegen deze verwarring van betekenissen. Zo zegt hij in zijn boek *Intentionality* (1983):

> "One of the most pervasive confusions in comtemporary philosophy is the mistaken belief that there is some close connection, perhaps even an identity, between intensionality-with-an-s and intentionality-with-a-t. Nothing could be further from the truth. They are not even remotely similar Intentionality-with-a-t is that property of the mind (brain) by which it is able to represent other things; intensionality-with-an-s is the failure of certain sentences, statements, etc., to satisfy certain logical tests for extensionality. The only connection between them is that some sentences about intentionality-with-a-t are intensional-with-an-s" (Searle 1983, 24) (7).

Ik zelf gebruik de term 'intentionaliteit' om te verwijzen naar een eigenschap van personen en hun psychologische toestanden waardoor ze op iets anders gericht kunnen zijn, iets anders kunnen representeren, en (met hun representaties) ergens naar kunnen verwijzen Met 'intentionele termen' bedoel ik termen die verwijzen naar psychologische of propositionele attitudes, zoals meningen en wensen enz 'Intensioneel' noem ik niet-extensionele zinnen, die Chisholm's logische kenmerken hebben Zinnen over intentionaliteit, die intentionele termen bevatten, zijn doorgaans intensioneel

De drie auteurs die ik zal behandelen gebruiken elk de term in een wat andere betekenis Ook hun spelling verschilt, en dat weer om verschillende redenen Om aan te geven hoe ze ieder precies de term gebruiken zal ik ze zelf aan het woord laten

M Boden.

> "As I use it, thought or behavior is "intensional" in that it is directed on a psychological object A man s intentions - and his intentional behavior - are intensional in this sense, so also are his hopes and fears, his knowledge and beliefs, his perceptions and illusions . the intensional object (the object of thought) can be described only by reference to the subject's thoughts, such as his purposes, beliefs, expectations, and desires There may be no actual thing with which the object of thought can be sensibly identified". (Boden 1972, 48-49)

Deze definitie van 'intensionality' blijft vrij dicht bij Brentano's begrip van intentionaliteit. Vervolgens wijdt Boden enige aandacht aan de logische kenmerken van intensionele zinnen a la Chisholm, en oppert ze de hypothese dat alle intensionele zinnen, ook de modale, te maken hebben met de gerichtheid van de geest op een object (Boden 1972, 62) Tenslotte legt ze in een voetnoot uit dat ze het woord met een 's' spelt om verwarring te voorkomen met het woord 'intentional' dat 'opzettelijk' betekent Het is niet haar bedoeling om de term uitsluitend te gebruiken als een logische eigenschap van zinnen Overigens draagt ze in diezelfde voetnoot waarin ze haar spelling en gebruik van de term uitlegt toch weer enigszins bij tot de algemene verwarring.

"When spelled with an *s*, the word more clearly suggests a contrast to the logician's term "extension" and thus emphasizes the logical peculiarities referred to in the text. When spelled with a *t*, the word more clearly suggests a psychological context and is particularly likely to bring to mind the psychological term "intention" " (Boden 1972, 370) (8).

J. Fodor houdt in zijn bundel *Representations* (1981a) geen consequente definitie van de term intentionaliteit aan. In de verschillende artikelen gebruikt hij verschillende betekenissen en verschillende spellingen. Zo zegt hij in een noot bij de Introduction.

"... I have generally favored the spelling "intens*i*onal" for the concept that's connected with opacity, reserving "intent-ional" for contexts connected with intent" (Fodor 1981a, 318)

Hier lijkt hij 'intensioneel' te gebruiken als eigenschap van zinnen, namelijk verwant met referentiële opaciteit (zie boven). Maar in een later artikel in de bundel, getiteld 'Three cheers for propositional attitudes', dat oorspronkelijk uit 1979 is maar ingrijpend gereviseerd is voor deze bundel, zegt hij:

".. I use 'intens*i*onal' to mean, in effect, *opaque*, and I use 'intentional' to mean *opaque and psychological*. Some intensional contexts (e.g., modal ones) are thus nonintentional; but not vice versa" (Fodor 1981a, 322).

De verwijzing naar 'opzettelijkheid' is hier helemaal verdwenen. En bij weer een ander artikel, 'Computation and reduction' uit 1978 stelt hij het volgende.

"When I speak of intensionality, I shall usually have two related facts in mind First, that psychological states (including specifically, propositional attitudes) are typically individuated by reference to their *content;* second, that

> expressions that occur in linguistic contexts subordinate to verbs of propositional attitude are typically nonreferential It is notoriously hard to say how, precisely, the first of these facts is to be construed or what precisely, the relation between the two facts is" (Fodor 1981a, 323-324).

Toch is in de context van zijn werk meestal duidelijk wat hij bedoelt met 'intentionality' en 'intensionality' nooit opzettelijkheid, en altijd, tenzij uitdrukkelijk het tegendeel vermeld wordt, zowel de referentiele opaakheid (van toeschrijvende zinnen) als het ergens op gericht zijn of inhoud hebben (van psychologische toestanden)

Dennett geeft in zijn bundel *Brainstorms* (1978) een eenduidige definitie van het begrip 'intentionaliteit'

> "For me, as for many recent authors, intentionality is primarily a feature of linguistic entities - idioms, contexts - and for my purposes here we can be satisfied that an idiom is intentional if substitution of codesignative terms do not preserve truth or if the "objects" of the idiom are not capturable in the usual way by quantifiers" (Dennett 1978b, 3).

Ofschoon hij 'intentionality' uitsluitend ziet als 'intensionaliteit', als eigenschap van linguïstische entiteiten in de zin van Chisholm, zegt hij toch Brentano's notie van intentionaliteit te gebruiken (Dennett 1978b, 3).

Al deze citaten zijn zo uitvoerig gegeven, omdat het begrip intentionaliteit een sleutelpositie inneemt in de oplossingen die de drie auteurs voorstellen voor het lichaam-geest probleem. Intentionaliteit is volgens allen zowel het kenmerk dat het psychische onderscheidt van het fysische, als het punt waarop de een of andere vorm van verzoening of reductie moet plaatsvinden. Dit geldt voor Boden, voor Fodor, en heel expliciet ook voor Dennett, die zegt:

> "The concept of an intentional system ... is made to bear a heavy load. It has been used here to form a bridge connecting the intentional domain (which includes our

'common sense' world of persons and actions, game theory, and the "neural signals" of the biologist) to the non-intentional domain of the physical sciences. That is a lot to expect of one concept, but nothing less than Brentano himself expected when, in a day of less fragmented science, he proposed intentionality as the mark that sunders the universe in the most fundamental way, dividing the mental from the physical" (Dennett 1978b, 22).

Alle auteurs willen dus het begrip intentionaliteit gebruiken als spil voor hun oplossing van het lichaam-geest probleem Zoals gezegd, ik zelf gebruik de term 'intentionaliteit' om te verwijzen naar een eigenschap van personen en hun psychologische toestanden, waardoor ze op iets gericht kunnen zijn, iets buiten zichzelf kunnen representeren voor zichzelf, waarbij dat iets niet of niet zo als in de representatie beschreven in de buitenwereld hoeft te bestaan 'Intensioneel' noem ik niet-extensionele zinnen die Chisholm's logische kenmerken hebben. En 'intentionele termen' zijn in mijn gebruik termen die verwijzen naar psychologische of propositionele attitudes. Wanneer ik andere auteurs citeer laat ik natuurlijk hun eigen spelling staan, wanneer ik zelf over ze spreek en hun positie beschrijf of bekritiseer, hanteer ik mijn eigen spelling. Dit houdt in dat ik zal proberen aan te tonen dat Boden het soms over intentionaliteit-met-een-t en soms over intensionaliteit-met-een-s heeft, en dat het onvoldoende uit elkaar houden van beide betekenissen haar positie op een fundamenteel punt onduidelijk maakt Fodor heeft het volgens mij vrijwel steeds over intentionaliteit-met-een-t En Dennett, ofschoon hij zegt Brentano's notie te gebruiken, heeft het vrijwel uitsluitend over intensionaliteit-met-een-s.

1.4.3.1. Boden's fysicalistische oplossing voor het lichaam-geest probleem.

In de cognitieve psychologie wordt het fysicalisme als vanzelfsprekend beschouwd Toch gebruikt men er, zoals gezegd, een mentalistisch vocabulaire. men spreekt er onbeschroomd van meningen en wensen en

doelstellingen En men is er, in reactie op het behaviorisme, van overtuigd dat het gedrag van de mens (of van andere cognitieve systemen) niet direct door stimuli uit de buitenwereld wordt bepaald.

Men ziet in de cognitieve psychologie evenwel geen tegenspraak tussen het mentalisme in de theorievorming en het fysicalisme als basisthese. Men denkt immers: er bestaan computers, die zijn net als mensen. We praten erover in mentalistische termen. Desondanks zijn computers tevens volledig fysische systemen waar volledig fysicalistische verklaringen voor bestaan. Zo ook zijn mensen volledig fysische systemen waar we zowel in mentalistische als in fysicalistische termen over kunnen praten. Op deze manier meent men het lichaam-geest probleem te hebben opgelost of onschadelijk gemaakt.

De filosofe/psychologe Margaret Boden is de eerste geweest (voor zover ik kan zien) die deze gedachtengang, zoals die min of meer impliciet leeft onder de cognitieve psychologen, expliciet heeft gemaakt (9), namelijk in haar artikel 'Intentionality and physical systems' (1970, herdrukt in Boden 1981) en haar boek *Purposive explanation in psychology* (1972).

Boden's redenering gaat als volgt (10): Het S-R paradigma van het behaviorisme is niet houdbaar. Het is de wereld zoals die bestaat voor een bepaald psychologisch subject, en niet de puur fysische stimulus, die een rol speelt in de verklaring van het gedrag. Men moet zich concentreren op de intentionele of subjectieve omgeving, en niet op de fysische omgeving, om het gedrag van een subject te verklaren (Boden 1972, 22). Daarom zijn in de psychologie mentalistische termen nodig die verwijzen naar wat het subject denkt en verwacht en wil van de wereld; en daarom moet het gedrag verklaard worden in intensionele zinnen, in zinnen met bepaalde logische kenmerken à la Chisholm (Boden verwijst expliciet naar Chisholm). Het gedrag van de mens wordt niet direct door de wereld bepaald, maar door hoe hij of zij zich de wereld voorstelt, door interne representaties van de wereld.

Ditzelfde geldt evenwel ook voor computers, aldus Boden. Ook hun gedrag wordt niet direct door de buitenwereld bepaald, maar door interne representaties. Ook hun gedrag moet beschreven worden in intentionele termen.

"Analogous logical features would characterize descriptions

> and explanations relating to artificial information-processing systems whose performance is controlled by such internal representations. A psychological being or subject is a physical system organized in this fashion The mind should be thought of as a set of such representational models, systematically interlinked in certain ways. Only if the concept of mind is interpreted in this sense can one understand how it is possible for the mind to act on the body" (Boden 1972, 3)

Wanneer we een psychologisch wezen - de mens - beschouwen als een fysisch systeem met interne representaties, dan kunnen we begrijpen hoe de 'geest' op het lichaam kan werken, aldus Boden, en dan hebben we het lichaam-geest probleem opgelost. De mens is een fysisch systeem met interne representaties, net als de computer. En een computer is - in zekere zin - een psychologisch wezen, net als de mens Het is gebruikelijk, maar ook nodig, over een computer te spreken in termen die ontleend zijn aan de gewone psychologie. "Hij probeert een bepaald doel te bereiken, ontmoet moeilijkheden, creeert een nieuw subdoel, verwacht dat dat gemakkelijker te bereiken valt" enz. Wanneer je zou weigeren om een geprogrammeerde computer in dit soort termen te beschrijven, zo zegt Boden, dan zou je de enige manier opgeven die er is om bepaalde belangrijke feiten over de structuur en werking van het programma uit te drukken (Boden 1972, 117).

Het bestaan van computers vormt een argument om aan te nemen dat intentionaliteit een eigenschap van zuiver fysische systemen kan zijn.

> "Cybernetic research ... can provide support for the reductionist view of intentionality, for the concept of an internal representation or model has found its way into the recent cybernetic literature, and the physical bases of models in machines are fully understood" (Boden 1972, 125).

In een artikel uit 1970 (herdrukt in een bundel uit 1981) geeft Boden een uitgewerkt voorbeeld van een machine die de kenmerken van intentionaliteit vertoont. Ze geeft dan aan hoe je een robot zou moeten

bouwen die de verschijnselen vertoont van hysterische verlamming Daarmee wil ze laten zien hoe een fysische of fysiologische verklaring in principe zo'n geval aankan, en hoe zo n verklaring ten grondslag kan liggen aan de intentionele kenmerken van gedrag (Boden 1981, 56).

De verschijnselen van hysterische verlamming zijn in verschillende opzichten vreemd Ze zijn niet het gevolg van enige duidelijke fysische beschadiging of verwonding En ze kunnen onder hypnose weer opgeheven worden Er is dus geen eenvoudige lichamelijke verklaring te geven voor zo'n verlamming in termen van een beschadigd motorneuron, of groep neuronen Maar het vreemdste van alles is wel dat de grenzen van de hysterische verlamming niet overeenkomen met enige anatomische grenzen van zenuwbanen De verlamde spieren komen niet overeen met enige groep van spieren die door een of meerdere zenuwen geinnerveerd worden De grenzen van de verlamming komen wel overeen met de lekenopvatting van anatomische grenzen Zo wordt in het dagelijks leven de arm gezien als een eenheid, begrensd door een lijn die overeenkomt met het armsgat van een mouwloos bloesje, en de hand als een eenheid, begrensd door een lijn rond de pols De verlamming betreft alle bewegingen, en geen andere, van het gebied dat de patient beschouwt als zijn hand Anatomisch gezien, gaan de zenuwen die die gebieden innerveren ook naar andere, onaangedane gebieden Maar de patient weet dat niet Zoals Boden opmerkt, het is zo iemand waarschijnlijk nooit opgevallen dat, als een deel van de hand 'slaapt', de pink en de aanliggende zijde van de ringvinger gevoelloos zijn, maar de andere kant van de ringvinger niet Voor hem is een vinger een vinger, en niet iets wat in tweeen gedeeld is De hysterische verlamming kan alleen verklaard worden als melding gemaakt wordt van niet-fysische gegevens, nl het concept van, of de gedachten en opvattingen over, een arm' of een 'hand' die de patient heeft

Toch is ook dit geval, waar duidelijk concepten en gedachten het lichamelijk gedrag besturen, geen ontkrachting van de grondthese van het fysicalisme, aldus Boden (1981, 58) Neem een computer, gebouwd als een robot, met vingers en tenen, ledematen die kunnen buigen en strekken net als bij de mens De bedrading van de robot loopt volgens hetzelfde plan als het menselijk zenuwstelsel Zo lopen de draden die

gaan naar de pink en de buitenkant van de ringvinger samen in dezelfde isolatiebuis. Vervolgens krijgt het hoofd van de robot foto-electrische cellen, en worden op de verschillende delen van zijn lichaam verschillende gekleurde lichtjes bevestigd. Zo kan hij zijn eigen lichamelijke bewegingen op een elementair niveau onderscheiden. Tenslotte installeren we in de robot een negatief *feedback* mechanisme dat bepaalde bewegingen onmiddellijk kan inhiberen.

Door middel van het systeem van gekleurde lichtjes kunnen we de robot functionele 'interne modellen', of 'concepten', geven van lichaamsdelen die corresponderen met de lekenopvattingen en niet met zijn anatomie. Voor de leek vormen vijf van de vingers en een palm samen de eenheid 'linkerhand', die begrensd wordt door het polsgewricht. Dus we laten alle gekleurde lichtjes die op die delen bevestigd zijn vallen onder een kopje in het geheugen van de robot En de inhibitie-instructie 'stop beweging' is geassocieerd met dat kopje, dat we 'linkerhand' noemen.

De robot, zo geconstrueerd, zou een hysterische verlamming van zijn linkerhand hebben (Natuurlijk gaat het hier om slechts één kenmerk van de hysterische verlamming, namelijk dat *common sense* opvattingen van lichaamsdelen en hun grenzen er een rol in spelen, en niet de anatomische grenzen van lichaamsdelen Alle andere aspecten van de hysterie, zoals de speciale betekenis die het geaffecteerde lichaamsdeel heeft voor de hystericus, blijven volledig buiten beschouwing). De verrichtingen van de robot kunnen alleen verklaard worden in termen van het 'model' of 'concept' van zijn linkerhand dat hij heeft, en niet in termen van de bedrading van de hand, zelfs al weten we elk mechanisch detail. Die détails kunnen door ingenieurs veranderd worden, maar de gehele structuur van het gedrag blijft hetzelfde· de hand blijft verlamd (Boden 1981, 62-63).

Met dit voorbeeld meent Boden te kunnen laten zien dat ook geheel fysische systemen gedrag kunnen vertonen dat intentionele kenmerken heeft Gedrag wordt, zo zegt ze, grotendeels gemedieerd en gecontroleerd door middel van interne - en vaak idiosyncratische - representaties van de omgeving, en niet direct door de omgeving. En dit feit ligt ten grondslag aan de intentionele kenmerken van gedrag, en aan de logische kenmerken van intensionele zinnen over zulk gedrag. Boden meent aldus een (fysicalistische) oplossing gevonden te

hebben voor het probleem hoe het mogelijk is dat de geest inwerkt op het lichaam

1 4 3.2 Kritiek op Boden's oplossing

Boden's oplossing voor het lichaam-geest probleem is op het eerste gezicht erg aantrekkelijk. Er blijft alle ruimte en noodzaak ook om te spreken over intentionaliteit, over hopen, geloven, denken, verwachten, willen enz Ze doet erg veel moeite om te laten zien dat haar fysicalisme niet ontmenselijkend is, en volstrekt verenigbaar met de zogenaamde humanistische psychologie Ze eindigt haar boek dan ook met de woorden.

> "To regard all purposive creatures as, basically, physical mechanisms is not to deny the reality of mind, nor to assert the inhumanity of man" (Boden 1972, 341).

Maar haar fysicalistische oplossing voor het lichaam-geest probleem laat nog wel veel vragen open. Haar positie heb ik gekarakteriseerd als voorbeeld van de grotendeels naïeve en impliciete opvattingen over het lichaam-geest probleem zoals die gangbaar zijn onder cognitieve psychologen Ze geeft een eerste explicitering van die opvattingen Ik wil nu aangeven op welke punten haar positie uitwerking behoeft.

Ten eerste is het nooit helemaal duidelijk in hoeverre en in welk opzicht volgens Boden computers net als mensen zijn. Soms gaat ze heel ver· de mens is een fysisch systeem met een organisatie als een computer (Boden 1972, 3). Dan weer meent ze dat computers hooguit in analoge zin net als mensen zijn, en spreekt ze over de computationele *metafoor* in de psychologie (b.v. Boden 1977, 121 en 1979, 49) (11) Haar robot uit het voorbeeld heeft een bedrading die overeenkomt met onze zenuwbanen (1981, 62-63), maar bij een andere gelegenheid benadrukt ze het feit dat het gedrag van een computer niet direct afhangt van de *hardware* van de machine. De vraag in hoeverre of in welk opzicht het fysisch systeem dat een computer is gelijk is aan het fysisch systeem dat de mens is wordt bij Boden niet echt gethematiseerd. Het is voor een belangrijk deel een empirische

vraag Boden redeneert zelf dat het bestaan van computers die net als mensen zijn ons nu eindelijk laat begrijpen hoe de geest op het lichaam kan werken. Maar dan moet wel aangetoond worden dat en in hoeverre computers net als mensen zijn

Ten tweede moet Boden aangeven wat de relatie is tussen de verklaringen in intentionele termen en de fysische verklaringen die beide van toepassing zijn op eenzelfde fysisch systeem Nu besteedt ze daar wel aandacht aan in haar boek uit 1972, maar haar uiteenzetting van die relatie is niet erg helder (zie ook Meijsing 1985) en heeft ook weinig navolging gevonden

Ten derde is bij Boden de status van een model of interne representatie niet duidelijk. Meestal lijkt het erop of een model expliciet aanwezig is in een systeem of organisme Boden spreekt dan over de fysische parameters van cerebrale modellen (1972, 309), en over een causaal mechanisme in de hersenen (1972, 307). En de robot uit het voorbeeld heeft 'interne modellen'. Maar soms is het of zo'n model of representatie alleen maar wordt toegeschreven op grond van het molaire gedrag, zonder dat er *in* het systeem of organisme iets expliciet aanwezig hoeft te zijn.

Dit onderscheid is van groot belang. Er is een verschil te zeggen dat een organisme of een systeem een interne representatie heeft, expliciet in zich heeft, of dat het zich gedraagt *alsof* het een representatie heeft Er is een verschil te zeggen dat een mens (of een machine) echt denkt en doelstellingen heeft, of dat je dat alleen maar, om pragmatische redenen, toeschrijft. Boden is hier niet echt duidelijk over. Zo schrijft ze:

> "Vocabulary drawn from everyday psychological language may be helpful in describing, explaining, and predicting machine performance" (Boden 1972, 137).

en:

> "To borrow a phrase from Laing, one's "initial intentional act" of choosing a linguistic level in describing behavior or performance determines the type of explanation that one will find appropriate" (Boden 1972, 134).

Het is niet duidelijk of Boden intentionaliteit wil toekennen aan

machines, wil beweren dat machines de eigenschap intentionaliteit hebben, of dat ze alleen in intensionele zinnen met intentionele termen over machines wil praten, omdat dat makkelijk is Om haar robot-voorbeeld te toetsen vraagt ze zich af hoe het haar robot vergaat met de logische criteria voor intensionaliteit van Chisholm (Boden 1981, 64). Hier wreekt zich mogelijk de onduidelijkheid over intentionaliteit als eigenschap van personen, en intensionaliteit als eigenschap van zinnen of taalgebruik. Beide mogelijkheden, de bewering dat sommige wezens intentionaliteit hebben, en de bewering dat het er alleen om gaat of je over sommige wezens in intensionele zinnen kunt praten, blijven open bij Boden (12).

Ten slotte is haar voorbeeld van de robot met de hysterische verlamming *question begging,* mede doordat ze intensionaliteit en intentionaliteit met elkaar verwart. Het moet een voorbeeld zijn van een fysisch systeem dat intentioneel gedrag vertoont. Daarbij gaat het erom dat het gedrag van de robot alleen verklaard kan worden onder verwijzing naar een bepaald intern model of eenvoudig 'concept'. Omdat de robot zo'n concept heeft, moet er in intensionele uitspraken over hem gepraat worden.

Maar heeft die robot wel een concept van een linkerhand? (Alle lichtjes op zijn lichaam zijn aanvankelijk gelijk voor hem. Alles wat de robot heeft zijn lichtgevoelige cellen in zijn hoofd). Door alle lichtjes die bevestigd zijn op vijf vingers en een palm te zetten onder het kopje 'linkerhand' heeft de robot zijn concept, volgens Boden.

Hier is evenwel van alles mis mee. Boden zegt letterlijk: Voor de leek vormen vijf van de vingers en een palm samen de eenheid linkerhand. Dat klopt al niet, het moeten de vijf vingers en de palm *van de linkerhand* zijn. Die beschrijving vóóronderstelt al het concept van de linkerhand, of op zijn minst het kunnen verwijzen naar de linkerhand. Maar zelfs als we dit punt triviaal vinden, dan is er een nog veel ernstiger bezwaar. Hoe kan de robot de lichtjes op een vinger onderscheiden van die op een teen? Alleen als hij al het concept van een vinger heeft! En Boden had juist uitgelegd dat ook de lekenopvatting van een vinger niet overeenkomt met enige anatomische eenheid. Het probleem herhaalt zich dan gewoon. Boden's robot is veel te simpel zoals hij nu beschreven is En om te zeggen: "We maken de robot net als onszelf, dan krijgt hij zijn concepten net als wij" is in

ieder geval *question begging:* we weten nog steeds niet goed hoe onze perceptuele en concept-vormende systemen in elkaar zitten (13).

Er is nog een andere mogelijkheid: we kunnen alle lichtjes die gemonteerd zijn op de hand een afwijkende kleur geven, bijvoorbeeld blauw. De instructie 'Stop beweging' wordt dan geassocieerd met blauw licht, en de robot moet een mechanisme ingebouwd krijgen dat alleen reageert op licht van een bepaalde golflengte. Maar waarom zou je dan nog zeggen dat de robot een concept heeft van een hand? Als toevallig een lichtje op zijn teen ook blauw is wordt die teen ook verlamd. En, nog veel ernstiger, waarom zou je dan nog spreken van een concept, of van intentionaliteit? De hysterische verlamming van de robot is nu niet afhankelijk van het concept van een hand, maar van licht van een bepaalde golflengte. Zijn bedrading moet dan zo zijn dat licht van een bepaalde golflengte bepaalde bewegingen inhibeert, namelijk de bewegingen van die delen waar de lichten aan bevestigd zijn. Weliswaar vormen die delen niet een eenheid volgens het (fysische) systeem van bedrading dat we anatomisch hebben genoemd, maar ze worden wel tot een eenheid via een ander systeem van bedrading en andere fysische kenmerken. Een feedback-mechanisme zonder enige concepten kan voor de verlamming zorgen. Zo'n robot heeft net zo veel of zo weinig intentionaliteit als een thermostaat. Natuurlijk is het mogelijk om in intensionele zinnen met intentionele termen te spreken over een thermostaat: "Hij probeert de temperatuur aangenaam te houden, je kunt hem laten denken dat het warm is door er een kaars onder te houden" enz. Sommige mensen praten ook zo over hun auto. Maar het gedrag van robot, thermostaat en auto kan ook volledig en adequaat beschreven worden in volledig extensionele zinnen.

Boden laat in het midden of het bij mensen en sommige machines noodzakelijk is om ze in intentionele termen te beschrijven, of dat het alleen maar mogelijk (en handig) is; ze laat in het midden of sommige organismen of systemen de eigenschap 'intentionaliteit' hebben, of dat je er alleen in intentionele termen over kunt praten, als dat zo uitkomt (14), en dat de zinnen die je dan gebruikt de eigenschap intensionaliteit hebben.

In de volgende hoofdstukken worden bovengenoemde vier punten uit Boden's positie nader uitgewerkt. De vraag in hoeverre en in welk opzicht mens en machine gelijk zijn wordt behandeld in hoofdstuk 2.

De relatie tussen verklaringen in intentionele termen en verklaringen in fysische termen wordt besproken in hoofdstuk 3, met name in 3.3.1. De positie dat intentionaliteit een eigenschap is van sommige organismen, die echt interne representaties hebben, en echt meningen en wensen enz., wordt uitgewerkt aan de hand van een analyse van het werk van Jerry Fodor En de positie dat het soms handig is om aan systemen interne representaties en meningen en wensen toe te schrijven ofschoon ze die niet echt hebben, en dat de zinnen die je voor die toeschrijvingen gebruikt intensioneel zijn, wordt behandeld aan de hand van een analyse van het werk van Daniel Dennett.

2 MENS EN COMPUTER IN DE COGNITIEVE PSYCHOLOGIE

2 1 Inleiding

We hebben aan de hand van een bespreking van het werk van Margaret Boden gezien da de cognitieve psychologie meent een fysicalistische oplossing te kunnen bieden voor het lichaam-geest probleem Men stelt mensen zijn zuiver fysische systemen, een soort machines Dit soort mens-machine vergelijkingen leek altijd heel hard en ontmenselijkend, maar is dat nu niet meer Immers, we hebben tegenwoordig een heel nieuw soort machine, de computer Het gedrag van zo'n computer kan in louter fysische termen beschreven en verklaard worden, maar kan ook beschreven en verklaard worden in termen die verwijzen naar psychologische toestanden, in intentionele termen Wanneer die twee soorten beschrijvingen en verklaringen voor een machine mogelijk zijn, waaraan verder niets mysterieus is, dan is het nu ook begrijpelijk dat voor de mens twee heel verschillende soorten beschrijvingen en verklaringen van toepassing zijn, de mens is ook een volledig fysisch systeem, net als de computer

We hebben ook gezien dat deze oplossing van het lichaam-geest probleem op twee poten steunt, een empirische en een apriori poot, die elk verder uitgewerkt dienen te worden In dit hoofdstuk wil ik de empirische poot uitwerken, de poot waarin gesteld wordt "Kijk maar, computers (en meer nog, robots) zijn echt net als mensen in cognitief opzicht, er is dus aan mensen niets mysterieus of geheimzinnigs in de zin van een lichaam-geest probleem" Allereerst zal ik kort uitleggen wat een computer is, en wat bedoeld wordt met de term Artificiele Intelligentie (AI) Een aantal richtingen binnen de AI zal worden onderscheiden

Vervolgens wordt onderzocht hoe sterk die empirische poot van de fysicalistische oplossing van het lichaam-geest probleem is er wordt onderzocht in hoeverre computers net als mensen zijn Hierbij komen twee discussies aan de orde die van het *grain*-probleem en die van het *frame*-probleem

In het *grain* probleem wordt de vraag gesteld met welke *grain*, welke korrel van oplossing men moet kijken om de mens-machine gelijkheid te

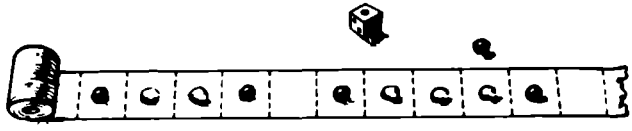
kunnen zien. Er wordt een historisch overzicht gegeven van het zoeken naar criteria voor die gelijkheid Ter sprake komen de perceptrons, de Turingtest, Pylyshyn's methodologische voorstellen voor equivalentie-criteria en een vergelijking tussen Fodor en Pylyshyn. Die vergelijking is voor mijn betoog van belang omdat Fodor het domein van de AI op een andere plaats legt dan Pylyshyn. Volgens hem is het domein van de AI beperkt Dit leidt tot de bespreking van het *frame*-probleem

In het *frame*-probleem wordt de vraag gesteld of er misschien menselijke cognitieve verrichtingen zijn die een computer niet kan uitvoeren. Ter sprake komen argumenten van Fodor en Dreyfus ten aanzien van het *frame*-probleem.

De conclusie van dit hoofdstuk zal luiden dat de empirische poot van de oplossing van het lichaam-geest probleem die de cognitieve psychologie meent te kunnen bieden helemaal niet empirisch is. Wanneer men op apriori gronden gelooft dat de mens een soort computer is, is het inzichtelijk dat in principe computers in de toekomst net als mensen zullen zijn. Gelooft men echter op apriori gronden dat de mens geen fysische machine is, of in ieder geval niet een soort computer, dan volgt daaruit dat computers nooit net als mensen zullen zijn. De bestaande geprogrammeerde computers zijn nog te onvolkomen om hetzij de ene hetzij de andere apriori overtuiging aangaande de mens empirische steun te verlenen.

2.2.1. Wat is een computer?

Een computer is, anders dan de naam doet vermoeden, niet alleen een rekenmachine. Een computer is een symbool-manipulator Hoewel de huidige computers buitengewoon indrukwekkend zijn, is het onderliggende principe vrij simpel. Het is belangrijk om te zien waar een computer zijn macht vandaan heeft.



Figuur 1 Begin van het spel (overgenomen uit Weizenbaum 1983, 64)

Als de dobbelsteen laat zien een·	*En de steen onder de aanwijzer is*	*Draai dan de dobbelsteen zo dat boven ligt*	*Vervang de steen door.*	*Verschuif de aanwijzer naar·*
1	geen	3	wit	links
1	zwart	2	geen steen	links
1	wit	1	wit	links
2	geen	2	geen steen	links
2	zwart	3	geen steen	links
2	wit	5	geen steen	rechts
3	geen	3	geen steen	links
3	zwart	4	geen steen	rechts
3	wit	5	geen steen	rechts
4	geen	4	geen steen	rechts
4	zwart	1	zwart	rechts
4	wit	6	wit	links
5	geen	5	geen steen	rechts
5	zwart	1	zwart	rechts
5	wit	1	wit	links
6	geen	0	geen steen	rechts
6	zwart	0	zwart	rechts
6	wit	3	wit	links

Regelnummer	*Symbool onder de koppen*	*Volgende regelnummer*	*Te schrijven symbool*	*Tape-bewegings-richting*
xIx	xOx	xIIIx	xIx	xIx
xIx	xxx	xIIx	xOx	xIx
xIx	xIx	xIx	xIx	xIx
xIIx	xOx	xIIx	xOx	xIx
xIIx	xxx	xIIIx	xOx	xIx
xIIx	xIx	xIIIIIx	xOx	xOx
xIIIx	xOx	xIIIx	xOx	xIx
xIIIx	xxx	xIIIIx	xOx	xOx
xIIIx	xIx	xIIIIIx	xOx	xOx
xIIIIx	xOx	xIIIIx	xOx	xOx
xIIIIx	xxx	xIx	xxx	xOx
xIIIIx	xIx	xIIIIIIx	xIx	xIx
xIIIIIx	xOx	xIIIIIx	xOx	xOx
xIIIIIx	xxx	xIx	xxx	xOx
xIIIIIx	xIx	xIx	xIx	xIx
xIIIIIIx	xOx	xx	xOx	xOx
xIIIIIIx	xxx	xx	xxx	xOx
xIIIIIIx	xIx	xIIIx	xIx	xIx

Tabel 1 De regels (overgenomen uit Weizenbaum 1983, 65 en 71)

Stel je het volgende spel voor (Weizenbaum 1976) We hebben nodig: een rol wc-papier, een heleboel witte steentjes, vijf zwarte stenen en een dobbelsteen. Begin met een situatie zoals aangegeven in figuur 1. De zwarte steen boven de rol noemen we de 'aanwijzer'. Volg dan zorgvuldig de regels uit tabel 1, dat wil zeggen, draai, volgens de regels, steeds de dobbelsteen op een nieuw cijfer, vervang de steen op het wc-papier, verplaats de aanwijzer en zoek dan de volgende regel op. Het spel stopt als de dobbelsteen op 0 gedraaid wordt. (N.B. het is geen leuk spel: de speler heeft geen enkele keus en de dobbelsteen wordt niet geworpen of voor kans gebruikt. Maar het is dan ook niet echt een spel om te spelen.)

Om nu wat gemakkelijker over het spel te kunnen spreken veranderen we de notatie: 'zwart' wordt 'X'; 'wit' wordt '1' en 'geen' wordt '0'. De aanvangsconfiguratie is dan ...000X11X0X11$\underline{1}$X00... De aanwijzer is alleen een geheugensteuntje; ik heb hier de gemarkeerde plaats onderstreept. Als we nu de rijtjes van '1'-en als nummers interpreteren - '111' betekent 3 - dan zien we dat de 'X' slechts als interpunctie dient. De eindstand van het spel is dan ...00$\underline{0}$11111X00... Deze configuratie kan gezien worden als de som van de twee getallen uit de beginstand. Het spel vormt een optelprocedure.

Het is niet moeilijk om te zien hoe een machine gebouwd kan worden die deze procedure uitvoert. We vervangen de rol wc-papier door een geluidsband en de speler door een bandrecorder. De regels moeten wel een beetje veranderd worden: nu beweegt de band en niet de aanwijzer - rechts en links moeten dus verwisseld worden. We installeren zes relays in de recorder om de informatie van de stand van de dobbelsteen in op te slaan. Verder nemen we een hoge toon voor een 'X', een middeltoon voor '1' en een lage toon voor '0'. Als we nu een band hebben opgenomen volgens de beginstand van het spel, en we zetten de gemarkeerde toon op de band onder de koppen, en we zorgen dat relay 1 gesloten is en alle andere relays open, dan kan het spel beginnen. De bedrading van de bandrecorder is zo dat bijvoorbeeld als relay 1 gesloten is en een hoge toon (X) gehoord wordt, de hoge toon wordt uitgewist en overspeeld met een lage toon (0), de band een stap naar rechts verschuift, relay 1 wordt geopend en relay 2 gesloten. Dit is een uitvoering van regel 2 uit tabel 1. We hebben nu een werkende optelmachine: de regels en de bouw van het

apparaat zorgen daarvoor.

We hebben maar drie verschillende symbolen op de band. Maar we kunnen alle regels ook in die drie symbolen uitdrukken. Laten we spreken van de 'staat' van de machine als zijnde het nummer van de dobbelsteen of van het gesloten relay. We kunnen nu de vijf delen van de regels in de volgende sequentie schrijven: huidige staat - symbool onder de kop - volgende staat - te schrijven symbool - richting van de bandbeweging. En we nemen de volgende code aan:

1) 'X' is een interpunctie

2) We geven een getal aan met het corresponderende aantal '1'-en

3) '0' staat in de context 'richting' voor links en '1' voor rechts.

We krijgen dan regeltabel 2. Ook als we alle regels zonder spatie achter elkaar doorschrijven hebben we een complete beschrijving van onze machine. Een machine van het soort van onze optelmachine, die een band heen en weer schuift met telkens één stap tegelijk, leest en schrijft op zo'n stukje band, in een andere staat overgaat enz., heet een *Turingmachine*. Een Turingmachine kan geheel beschreven worden in termen van een verzameling vijftallen in de vorm: huidige staat - te lezen symbool - volgende staat - te schrijven symbool - bewegingsrichting van de band. Die vijftallen kunnen op hun beurt geschreven worden in de notatie die de Turingmachine ook als input op de band accepteert.

De Engelse wiskundige Alan Turing heeft in 1936 het volgende bewezen: "Er bestaat een Turingmachine U, met als alfabet de symbolen '1' en '0', zodanig dat, gegeven elke procedure geschreven in elke preciese en niet-ambigue taal, en gegeven een Turingmachine L die de transformatieregels van die taal belichaamt, de Turingmachine U de Turingmachine L kan imiteren in de uitvoering van die procedure". Dus in ons geval is er een Turingmachine U die als input accepteert een band met daarop zowel de beginconfiguratie als de machinebeschrijving van onze optelmachine, en die dan de optelmachine *imiteert*. En de universele Turingmachine U kan *alle* Turingmachines op die manier imiteren. Een universele Turingmachine is eigenlijk geen echte machine maar een wiskundige specificatie van een machine. Een Turingmachine, als wiskundige specificatie, heeft een oneindige band. Een werkelijke machine, zoals de bovenbeschreven bandrecorder, is natuurlijk eindig, en heeft een eindige band. Moderne computers lijken

nauwelijks nog op de machine die Turing beschreef. Vele hebben bijvoorbeeld de mogelijkheid om een heleboel banden tegelijk te manipuleren, en belangrijker nog, de meeste hebben zeer grote informatie-opslagruimtes, die functioneel gelijk zijn aan een set relays die elk gesloten (aan) of open (uit) kunnen zijn (zogenaamde flip-flops). Een set van tien zulke relays kan 1024 verschillende staten aannemen. Het is niet ongewoon voor een middelgrote computer om meer dan een miljoen flip-flops te hebben, terwijl de chip nog meer mogelijkheden voor informatie-opslag biedt. Toch, ondanks die verschillen in omvang en complexiteit, is in principe iedere moderne computer, de *special purpose* machines (dat wil zeggen machines die gebouwd zijn om één nauw omschreven taak uit te voeren) uitgesloten, een universele Turingmachine (behoudens geheugenbeperkingen). En dat betekent dat in principe iedere moderne computer iedere andere computer kan imiteren (15).

De vraag rest nu nog: voor welke procedures kan men een Turingmachine specificeren, dus een machine imiteerbaar door de universele Turingmachine, dus een machine imiteerbaar door een moderne computer? Turing (1936) antwoordde op die vraag: voor elk proces dat men een effectieve procedure kan noemen. En een effectieve procedure is een verzameling regels die een speler van moment tot moment exact vertellen wat te doen. Let wel: de speler (of de computer) die een effectieve procedure uitvoert 'weet' en kan in principe alleen wat *iedere regel* hem vertelt te doen. Wat de procedure inhoudt is iets heel anders. Het was immers niet direct inzichtelijk dat het spel een optelprocedure was?

Een computer is geen rekenmachine, het is een symboolschuiver. We hebben gezien dat met behulp van het schuiven van symbolen een rekenkundige procedure geïmplementeerd is: optellen. Voor de optelmachine hebben we een notatie ontwikkeld waarin we die machine kunnen beschrijven. Het alfabet bestond uit drie symbolen: 'X', '0' en '1'. Natuurlijk zouden rijtjes in die symbolen betekenisloos blijven als we niet konden zeggen hoe ze geïnterpreteerd moeten worden. Maar zoals we zagen, we hebben ook een machine die als input neemt: de beschrijving van de eerste machine èn de input van de eerste machine. Deze Turingmachine voert de procedure van de eerste machine uit. Je kunt nu spreken van een 'machinetaal'. Wij interpreteren die taal: '0'

betekent links en '1' rechts enz Maar de machine hoeft die taal niet te interpreteren (of in een andere taal om te zetten) bij 0 gaat de band gewoon naar links en bij 1 naar rechts - de machine is zo gebouwd Ieder symbool in de machinetaal *veroorzaakt* een machinebeweging

Het begrip 'optellen' komt in die taal niet voor. Dat is een interpretatie die *wij* aan het gebeuren geven Met behulp van dat symboolschuiven kunnen we allerlei procedures opbouwen. met behulp van optellen definièren we vermenigvuldigen en aftrekken, met behulp daarvan definieren we machtsverheffen en worteltrekken enz Zo kunnen, hierarchisch, programma's opgebouwd worden. een zogenaamde routine voor het oplossen van vierkantsvergelijkingen maakt gebruik van onder andere een subroutine voor worteltrekken die gebruik maakt van een subroutine voor vermenigvuldigen die gebruik maakt van een subroutine voor optellen En optellen is een primitieve operatie in onze machine.

Het alfabet van alle moderne machinetalen bestaat uit twee symbolen· '0' en '1'. De vocabulaires en transformatieregels lopen uiteen, afhankelijk van hoe de machine gebouwd is. Het primitieve vocabulaire dat een programmeur kan gebruiken wordt bepaald door de operaties die in de machine ingebouwd zijn. In onze machine was optellen ingebouwd; in een andere kan worteltrekken ingebouwd zijn.

Nu is het zo dat programmeurs niet werken in machinetaal. Het maken van grote programma's in rijen '0'-en en '1'-en is gewoonweg ondoenlijk. We hebben gezien dat de operatie 'optellen' te vertalen is in machinetaal. Het is gemakkelijker om een programma te schrijven waarin de primitieve operaties in woorden worden uitgedrukt. Het is ook mogelijk om een programma te schrijven dat die woorden in hun juiste samenhang vertaalt in machinetaal We hebben dan een programmeertaal op een niveau hoger dan de machinetaal, de zogenaamde *assemblytaal.* Maar ook dat is nog heel vervelend programmeren: beschrijf maar eens het oplossen van een vierkantsvergelijking uitsluitend in termen van optellen.

Programmeurs hebben zich al gauw gerealiseerd dat de symbool-manipulerende capaciteiten van computers gebruikt konden worden om talen te vertalen op een nog hoger niveau. Maar dan komen er problemen. Een assemblytaal voor een bepaalde machine is een heel getrouwe weergave van de machinetaal (ofschoon in de machinetaal

helemaal geen begrippen uit de rekenkunde voorkomen, zie boven) Maar een hogere programmeertaal lijkt veel meer op natuurlijke taal - dat is ook het gemak ervan. Toch is het een formele taal en de rijkdom aan connotaties en ambiguïteit van de in de programmeertaal gebruikte termen is slechts schijn. De programmeertaal is eenduidig en precies, en het is de ontwerper van de taal die beslist welke betekenisaspecten van de gebruikte termen naar beneden worden vertaald. Bijvoorbeeld, in de programmeertaal lijkt de uitdrukking "Amsterdam is een stad" te gaan over Amsterdam, en over een plaats waar veel mensen bij elkaar wonen, mogelijk over uitgaansleven, junkies, rondvaartboten, Artis en nog veel meer In de machinetaal vertaald gaat die uitdrukking alleen over het feit dat een bepaald symbool een bepaalde relatie heeft met een ander symbool. De gebruikers van die programmeertaal komen vaak nauwelijks in aanraking met de vertaler; die kunnen niet weten welke betekenisaspecten uit de programmeertaal vertaald worden Dit kan wel eens misverstanden geven over wat het precies is dat een programma doet, bijvoorbeeld wanneer een programma-ontwerper zijn routines wat al te bloemrijke namen geeft waar het in feite om een zeer simpele operatie gaat (zie ook 4.6.1). Zoals Boden zegt:

> It is partly because a programmer can irresponsibly or over-enthusiastically define a basically trivial computational process under a superficially impressive label like 'FINDANALOG' or 'AESTHETIC SENSE', that one may be sceptical of a programmer's claims to have modelled analogical or aesthetic thinking within a program intended to do just this" (Boden 1981, 10; zie ook McDermott 1976).

2.2.2 Wat is AI?

Computergebruik speelt een rol in de cognitieve psychologie. In de eerste plaats kan de computer gezien worden als een werktuig bij de theorieconstructie, in de tweede plaats voor allerlei vormen van artificiële intelligentie. Wanneer gedrag beschouwd wordt als een reeks directe responsen op stimuli, dan kan een theorie van gedrag vrij

eenvoudig zijn. Men hoeft alleen de relaties tussen de stimuli en de responsen weer te geven. Wanneer die relatie geen lineaire functie vormt, moet men een *interveniërende variabele* postuleren, maar ook dat maakt de theorie niet bijzonder ingewikkeld. Anders wordt het als men gedrag ziet als het resultaat van een proces waarin naast de stimuli allerlei componenten en subsystemen binnen het organisme een bijdrage leveren. Wanneer er veel van zulke componenten gepostuleerd worden, kan het erg ingewikkeld worden om te specificeren hoe al die componenten met elkaar en met de inkomende stimuli interacteren om het gedrag te produceren. Hier kan de computer uitkomst bieden, als werktuig waarmee onafhankelijk opgestelde theorieën getoetst kunnen worden op volledigheid en consistentie. Zo stellen Anderson en Bower, bekende computergebruikers in de psychologie:

> "The computer is only a *computational tool* for explicitly checking the predictions of the theory, for determining whether all the specified mental processes are in fact fully specified and whether they can work together as claimed" (Anderson en Bower 1973, 143).

Programmeren is dan een vorm van expliciteren, en daarmee ook een soort test voor begrip Daarmee lijkt het op schrijven; programmeren is immers een vorm van schrijven. Vaak wanneer we menen iets te begrijpen en we gaan het opschrijven, blijkt dat we het helemaal niet zo goed begrijpen. We beginnen een zin met 'klaarblijkelijk' en zien opeens dat het helemaal niet zo klaar blijkt. Of we schrijven 'dus' en zien opeens dat onze redenering niet sluitend is. Maar vaak ook verhult de enorme rijkdom, flexibiliteit en ambiguiteit van onze taal dat soort feilen. Een computer laat zich echter niet door fraai taalgebruik misleiden, en heeft geen boodschap aan woorden als 'dus' en 'klaarblijkelijk'. Wanneer er ook maar iets impliciet blijft of ambigu, dan loopt het programma mis. En alle consequenties van onze theorie, niet alleen de bedoelde en gewenste, maar ook de onbedoelde, ongewenste, absurde consequenties worden meedogenloos duidelijk.

Zo heeft de opkomst van de moderne computer de ingewikkelde theorieconstructie van de cognitiewetenschap mogelijk gemaakt. Maar dit soort computergebruik is niet specifiek voor de cognitieve

psychologie In vele wetenschapsgebieden wordt de computer zo gebruikt Specifiek voor de cognitieve psychologie is het gebruik van computers voor de *artificiële intelligentie, AI.* 'AI' is de naam van een discipline of een tak van wetenschap, waar men zich bezighoudt met het ontwerpen van machines die taken kunnen verrichten welke, als mensen ze zouden verrichten, intelligentie zouden vereisen, een discipline dus die artificiele intelligentie als haar object heeft Het object heeft de discipline haar naam gegeven, 'AI-logie of 'AI-techniek' hadden als namen misschien meer voor de hand gelegen, maar de loop van de geschiedenis heeft het anders gewild.

Alan Turing voorzag in 1950 al de mogelijkheden (en moeilijkheden) van computers om intelligentievereisende taken te verrichten, in zijn artikel 'Computing machinery and intelligence' Hij eindigt zijn artikel met een voorstel voor toekomstig werk

> "We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult question Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English" (Turing 1950, 30)

Turing geeft niet aan hoe zulke programma's geschreven moeten worden, maar hij geeft wel drie belangrijke gebieden aan, waarin vele programma's geschreven zullen worden (schaak)spelen, patroonherkenning en taal Schaken werd het eerst aangepakt· in hetzelfde jaar, 1950, schrijft Claude Shannon, de man van de informatietheorie, een artikel over een schaakcomputer Ook Shannon schrijft geen schaakprogramma, maar in de tweede helft van de vijftiger jaren zijn er spelende computers (Newell et al 1963) Er werd gepoogd een vertaalmachine te construeren In 1954 bouwde Anthony Oettinger de eerste mechanische dictionaire In 1955 verscheen de eerste publicatie over patroonherkenning van Selfridge

Ook aan een algemene theorie over intelligentie werd gewerkt. In 1957 publiceren Newell, Shaw en Simon een klassiek geworden artikel

over de *Logic Theory Machine*, een machine die stellingen uit de symbolische logica kan bewijzen Het bijzondere van deze machine is dat hij niet met 'brute kracht' alle mogelijkheden uitprobeert, maar, net als mensen, gebruik maakt van een soort 'intelligente' vuistregels, die niet universeel correct zijn, maar vaak in minder tijd tot een oplossing leiden. Dit soort vuistregels worden, in navolging van George Polya (1954), 'heuristieken' of 'heuristische regels' genoemd Het idee van heuristische regels wordt door Newell, Shaw en Simon verder uitgewerkt in het project van de *General Problem Solver* (GPS), waarvan het eerste rapport verschijnt in 1960, en een afsluitend rapport in 1967 (Ernst en Newell). Een nieuw onderzoeksveld was ontstaan. de Artificiële Intelligentie (AI).

2 2.2 1. Artificiële intelligentie en cognitieve simulatie.

Traditioneel wordt er in de cognitiewetenschap een onderscheid gemaakt tussen twee soorten 'intelligente' systemen: diegene die in de eerste plaats ontworpen zijn om moeilijke taken te verrichten met behulp van wat voor slimme technieken er maar voorhanden zijn, en diegene die in de eerste plaats zijn ontworpen om menselijke cognitieve processen te simuleren. Het onderzoeksgebied dat zich bezighoudt met het construeren van de eerste soort systemen noemt men artificiële intelligentie in engere zin (16), de bezigheid die de tweede soort tracht te construeren noemt men cognitieve simulatie (b.v. Dreyfus 1979, Pylyshyn 1978).In plaats van 'artificiele intelligentie' versus 'cognitieve simulatie' gebruikt Weizenbaum (1976) de termen *performance mode* versus *simulation mode*. Hij spreekt ook over een *theory mode*, daarmee doelt hij op het gebruik van computers bij het toetsen van theorieën op volledigheid en interne consistentie

De beste manier om het onderscheid tussen artificiële intelligentie en cognitieve simulatie duidelijk te maken is misschien wel door middel van een vergelijking met het vliegen Vrijwel alle vroege pogingen om het vliegen te begrijpen of vliegende modellen te bouwen waren gebaseerd op het imiteren van het vliegen van vogels. Men kan aannemen dat de mythe van Ikaros, de Griekse held die vloog met vleugels die met was bevestigd waren en neerstortte toen hij te dicht bij de zon kwam

waardoor de was wegsmolt, het menselijk onvermogen weergaf om de vogels te imiteren. (Het hield natuurlijk ook een waarschuwing tegen *hubris* in. Misschien iets voor hedendaagse computerprogrammeurs?) Ook Leonardo da Vinci probeerde vleugels te ontwerpen om mee te vliegen als een vogel. Ook hij probeerde het vliegen van vogels te simuleren. Hij zei daarover dat een vogel een instrument is dat volgens natuurwetten werkt, en dat het binnen de vermogens van de mens ligt om dat instrument met al zijn bewegingen na te maken (zie Armer 1963). In later tijden lag het zwaartepunt meer op het begrijpen van die natuurwetten - van de aerodynamica - dan op het imiteren van vogels. Rond het midden van de vorige eeuw hielden mensen als Henson en Stringfellow, en wat later Langley, zich bezig met 'artificieel' vliegen. Zij achtten het hun taak vliegmachines te bouwen gebaseerd op wat voor werkende principes dan ook. En de historische vluchten van de gebroeders Wright in 1903 waren niet gebaseerd op de imitatie van vogels maar op algemene aerodynamische principes.

Artificiële intelligentie is dus een tak van wetenschap (of van techniek), die een eigen respectabiliteit heeft; ze hoeft zich in principe niets gelegen te laten liggen aan wat filosofen of psychologen in de loop der tijd beweerd of ontkend hebben over intelligentie of over machines. Als de opdrachtgever van een artificiele-intelligentie-onderzoek een automobielfabrikant is die een robot voor de assemblagelijn wil, of een speelgoedfabrikant die een schaakcomputer wil, dan gaat het erom dat de geleverde machines werken. Of ze dan net als mensen werken is niet van belang. Zo ook is het bij vliegtuigen van belang dat ze vliegen, niet dat ze net als vogels vliegen. In dit opzicht is de AI louter een toegepaste wetenschap of techniek.

Nu is het evenwel zo dat het onderscheid tussen beide soorten van programmeren, artificiele intelligentie en cognitieve simulatie, in de praktijk niet erg scherp is, daar is het een gradueel verschil in belangstelling en gerichtheid van de programmamakers. In de programma's is vaak een onderscheid te zien in generaliseerbaarheid en *power*. Vaak is er een omgekeerde evenredigheid tussen de breedte of generaliseerbaarheid van een methode (het gamma van taken waarop de methode van toepassing is) en de *power* van die methode (hoe goed hij het doet bij taken waarop hij van toepassing is) (Newell 1969). Veel paradepaardjes van de cognitiewetenschap zijn voorbeelden van

artificiele intelligentie in engere zin; deze programma's werken juist zo goed omdat ze zijn toegesneden op een zeer specifiek, nauw omschreven probleem. Dat geldt in het bijzonder voor zulke *powerful* programma's als expertsystemen en Winograd's beroemde SHRDLU (Winograd 1971) (17). De oplossingen voor de aan de programma's gestelde problemen zijn niet generaliseerbaar en weinig plausibel als psychologische modellen

Dit is een punt waar Dreyfus (1978) en Haugeland (1978b), belangrijke critici van de cognitiewetenschap, op hameren: de generaliseerbare en enigszins plausibele programma's zijn weinig indrukwekkend, en de indrukwekkende zijn niet generaliseerbaar en psychologisch niet plausibel. Dit punt wordt, als het om voorbeelden gaat die genoemd worden, niet betwist door cognitiewetenschappers. Men is het evenwel veelal oneens met Dreyfus' en ook wel Haugeland's diagnose, volgens welke dit probleem er een symptoom van is dat de cognitieve wetenschap op een dood spoor zit.

Van verschillende zijden is beweerd dat het onderscheid tussen artificiële intelligentie en cognitieve simulatie te verwaarlozen is. Pessimisten en tegenstanders van computergebruik (ik spreek hier niet van optimisme of pessimisme over de verrichtingen van computers, maar over de relevantie van computergebruik voor de cognitiewetenschap, de psychologie en de theorie van het mentale) menen dat voor het toetsen van psychologische theorieën echt programmeren niet nodig is. Het grootste deel van de programmeerarbeid gaat immers niet zitten in het specificeren van de theorie, maar in het vinden van oplossingen voor puur programmeertechnische problemen. Deze oplossingen zijn vanuit de theorie bezien volstrekt *ad hoc*, zodat ieder werkend programma in feite nauwelijks generaliseerbaar is, en veeleer tot de artificiële intelligentie behoort dan dat er nog sprake is van cognitieve simulatie. Bovendien is er nog een gevaar. In een enigszins uitgewerkt programma zijn de theorie-onderdelen en de programmeertechnische onderdelen niet van elkaar te onderscheiden. Het is mogelijk dat, wanneer men fouten, zogenaamde *bugs*, uit het programma haalt, er ongemerkt wijzigingen worden aangebracht in de theorie-onderdelen. Niet alleen zijn dan de *toevoegingen* aan de eigenlijke theorie in het programma volkomen *ad hoc*, er worden ook *ad hoc wijzigingen* in de

theorie aangebracht, die bovendien niet altijd voldoende gedocumenteerd worden De verleiding om een programma te laten werken kan erg groot zijn Weizenbaum (1976) schetst een overtuigend beeld van de compulsieve programmeur die altijd bezig is zijn programma's te verbeteren Zijn opwinding bereikt een koortsachtige hoogte wanneer hij een bijzonder hardnekkige *bug* op het spoor is Mocht op dat moment zijn computertijd bijna op zijn, dan kan hij op het laatste moment steeds grotere veranderingen aanbrengen, die hij niet of nauwelijks bijhoudt of aantekent Zo kan hij in enkele seconden roekeloos het werk van weken of zelfs maanden onherroepelijk vernietigen Natuurlijk zijn respectabele cognitiewetenschappers (meestal) geen compulsieve programmeurs, maar ook hun wordt de verleiding van een werkend programma soms teveel

Optimisten en voorstanders van computergebruik menen dat juist alle vormen van 'intelligente' systemen relevant zijn voor de cognitieve psychologie, zelfs programma s in zuivere artificiele-intelligentie-stijl Zoals zowel de vogel als de vliegmachine kunnen vliegen op grond van dezelfde aerodynamische principes, zo zegt men, zo kunnen ook zowel de mens als de machine intelligent gedrag vertonen op grond van dezelfde cognitieve principes Immers, het soort problemen waar artificiele-intelligentie-programma's een oplossing voor zoeken is duidelijk psychologisch Bijvoorbeeld, het identificeren van bepaalde klassen van patronen kan triviaal zijn Het is niet een probleem voor de cognitiewetenschap om bepaalde auditieve frequenties te laten herkennen, of bepaalde optische gradienten, of bepaalde golflengtecombinaties Zolang de equivalentieklasse van zo'n patroon eenvoudige, fysische kenmerken heeft, is het probleem van patroonherkenning niet een AI-probleem Wel een probleem voor de cognitiewetenschap is, een manier te bedenken om equivalentieklassen te herkennen die gedefinieerd zijn met behulp van psychologische criteria: klassen zoals de visuele patronen die corresponderen met iemands gezicht, of met de aanwezigheid van gewone voorwerpen op een foto, de auditieve patronen die corresponderen met een gesproken woord Dit zijn allemaal equivalentieklassen van fysische stimuli waarvoor een eenvoudige psychologische beschrijving bestaat, maar geen eenvoudige fysische beschrijving (Pylyshyn 1978) Mensen zien de visuele wereld niet in termen van optische gradienten en

golflengtecombinaties, maar in termen van bijvoorbeeld gezichten en voorwerpen; zij horen de auditieve wereld niet in termen van frequenties, maar in termen van bijvoorbeeld gesproken woorden. Als een computer de wereld ook zo kan klassificeren, dan is in ieder geval een manier aangetoond waarop het mogelijk is dat enig systeem zulke patronen kan herkennen. Als een computer de wereld ook zo kan klassificeren, hoe hij dan ook geprogrammeerd mag zijn, dan heeft hij toch iets gemeen met een mens In dit verband noemt de filosoof Daniel Dennett (1978b, 117) de AI, alle AI, een soort *gedachten-experimentele epistemologie*. AI houdt dan het midden tussen cognitieve psychologie en klassieke epistemologie. De cognitieve psychologie vraagt: "Hoe verricht dit systeem (de mens) de kennis- of intelligentievereisende taak X (of Y of Z)?" en probeert een antwoord te vinden door de mens te bestuderen. De epistemologie vraagt: "Hoe is kennis uberhaupt mogelijk?" en zoekt met aprioristisch redeneren naar *algemene* voorwaarden voor kennis. De AI vraagt: "Hoe kan enig systeem de kennis- of intelligentievereisende taak X (of Y of Z) verrichten?" en zoekt een antwoord door een *bepaald* systeem te ontwerpen dat X kan doen. Zowel de cognitieve simulatie als de artificiële intelligentie hebben de psychologie en de epistemologie wat te zeggen; in het eerste geval kunnen concrete, empirische hypotheses worden getoetst, in het tweede geval kan men een aantal meer abstracte principes aflezen van het model. Dat aflezen van abstracte principes is overigens een bijzonder moeilijke zaak: het is maar al te voor de hand liggend om de specifieke eigenschappen van deze schakende computer te verwarren met eigenschappen die ieder systeem moet hebben om te kunnen schaken.

2.2.2.2. Sterke AI en zwakke AI

Wanneer we niet al te pessimistisch zijn over het gebruik van computers in de psychologie anders dan als rekenmachines dan moeten we stellen dat de AI de psychologie en de epistemologie (en misschien ook de *philosophy of mind*) wat te zeggen heeft De vraag is nu: wat zegt de AI precies? De filosoof John Searle (1980) heeft een onderscheid gemaakt tussen zwakke AI en sterke AI. Ik wil dit

onderscheid overnemen en nog wat aanscherpen Het onderscheid tussen zwakke AI en sterke AI zoals ik het gebruik staat loodrecht op het onderscheid 'artificiele intelligentie - cognitieve simulatie' Volgens de zwakke AI is de voornaamste waarde van de computer bij het bestuderen van de geest, het mentale, dat hij een zeer krachtig werktuig is. De computer dwingt ons om hypothesen zeer rigoureus en precies te formuleren. Bovendien kunnen computersimulaties van bepaalde cognitieve processen zeer verhelderend zijn. Zowel de meer empirische, 'psychologische' simulatieprogramma's als de meer abstracte, artificiële-intelligentieprogramma's, kunnen veel licht werpen op de werking van mentale processen en, niet minder, op de structuur van de intelligentievereisende taken. De zwakke AI ziet de computer als een verhelderende metafoor voor de mens. Zo zegt Boden:

> "Computational theories of the mind are no more than that: theories One would not ask of a chemical theory that it fizz if put into a test-tube. Why, then, should one demand of a psychological theory that it see, or feel, if put into a computer? Psychologists try to understand human action and experience, not to mimic it ..." (Boden 1981, 49).

Maar volgens de sterke AI is de computer niet enkel een werktuig bij de bestudering van het mentale, niet enkel een verhelderende metafoor voor de mens. Het is inderdaad waar dat een chemische theorie niet bruist als hij in een computer wordt gestopt, maar een theorie van probleemoplossen lost, in een computer gestopt, wel degelijk problemen op De juist geprogrammeerde computer is niet een metafoor voor het mentale, hij *is* een 'geest', in die zin dat computers met de juiste programma's letterlijk kunnen begrijpen en andere mentale toestanden hebben. Zoals Pylyshyn zegt:

> "Given that computation and cognition can be viewed in the same common abstract terms, there is no reason why computation ought to be treated as merely a metaphor for cognition .. " (Pylyshyn 1980, 114).

Het onderscheid tussen zwakke en sterke AI staat, zoals gezegd,

loodrecht op het onderscheid tussen artificiele intelligentie en cognitieve simulatie. Volgens de zwakke AI zijn zowel de produkten van artificiële intelligentie als van cognitieve simulatie "niet het echte werk". Ze geven inzicht in de processen die een rol spelen bij het vervullen van bepaalde cognitieve taken, waarbij die processen in het geval van cognitieve simulatie wat meer aansluiten bij hoe mensen die taken verrichten, en in het geval van artificiële intelligentie wat meer taakgericht zijn. Maar volgens de zwakke AI verrichten computers niet echt die cognitieve taken; ze spelen niet echt schaak, ze voeren niet echt een conversatie, ze begrijpen niet echt, ze hebben geen echte mentale toestanden. Ze vormen een procesbeschrijving van het mentale zonder zelf mentaal te zijn, zoals een procesbeschrijving van een chemische reactie zelf geen chemische reactie is.

Volgens de sterke AI zijn de produkten van zowel artificiële intelligentie als van cognitieve simulatie wel "het echte werk"; ze spelen wel echt schaak, voeren echte conversaties, begrijpen echt, hebben echte mentale toestanden. Ze zijn zelf echt mentaal, zoals een synthetische chemische stof wel echt die chemische stof is.

Zwakke AI is een positie die nauwelijks controversieel is, en bovendien al gebleken is zeer vruchtbaar te zijn. De volgende paragrafen gaan zich bezighouden met sterke AI; met de AI die de computermetafoor letterlijk neemt, die stelt dat de computer gelijk is aan de mens(elijke geest). Daarbij zal in de eerste plaats onderzocht worden wat die claim precies inhoudt. Wat betekent het te zeggen: "De mens is gelijk aan de machine"? Daarna zal gekeken worden of die claim, in de meest recente, uitgewerkt versie, houdbaar is.

2 3. De mens-machine gelijkheid. Het grain-probleem.

Tegenstanders van sterke AI wordt nogal eens verweten dat ze flauw of oneerlijk zijn wanneer ze beweren dat computers niet gelijk zijn aan mensen. Die tegenstanders noemen dan een aantal verschillen op tussen computers en mensen: computers zijn niet van vlees en bloed, ze zijn gemaakt en niet geboren, ze leven niet, ze werken serieel en niet parallel, enz. Vanuit de sterke AI vindt men, zoals gezegd, deze tegenwerpingen flauw of oneerlijk, omdat er hier sprake is van de

ongelijkheid van irrelevante aspecten Maar dit argument kan ook omgekeerd worden Misschien zijn de aspecten waarop zij *gelijkheid* zien wel irrelevant of triviaal.

Gelijkheid en ongelijkheid zijn uiterst glibberige begrippen Tussen ieder paar entiteiten bestaat altijd zowel gelijkheid als ongelijkheid Het ligt er maar aan in welk opzicht men die gelijkheid of ongelijkheid wil zien Een mens is gelijk aan een steen ze hebben beide in vrije val nabij het aardoppervlak een valversnelling van ongeveer 10 m/s De leden van een een-eiige tweeling zijn ongelijk ze nemen te allen tijde een verschillende locatie in de ruimte in Wanneer niet gespecificeerd wordt in welk opzicht de gelijkheid of ongelijkheid bedoeld wordt, wordt er helemaal niets gezegd met de uitspraken "De mens is gelijk aan de computer" of "De mens is ongelijk aan de computer" En aangezien in ons alledaags, pretheoretisch taalgebruik de ongelijkheid van mens en computer meer voor de hand ligt dan de gelijkheid, lijkt het de taak van de sterke AI om te specificeren in welk opzicht de mens gelijk is aan de computer

Dit probleem van de specificatie van de gelijkheid van mens en computer heeft in de loop der jaren de nodige aandacht gekregen, en staat bekend onder de naam van het *grain*-probleem Het is het probleem met welke 'korrel van oplossing' je moet kijken om de gelijkheid tussen mens en machine te zien De extremen staan hierbij vast Met de fijnste korrel zie je de materialen waarvan beide zijn gemaakt. de mens voornamelijk uit koolstof-waterstof verbindingen, de computer voornamelijk uit metaal-siliconen verbindingen. In dit opzicht, met deze korrel, is er in ieder geval ongelijkheid Met de grofste korrel zie je alleen de input-output relaties Bij een geslaagd AI programma zie je dan, althans op een beperkt terrein van gedrag, gelijkheid tussen mens en computer. De computer is juist geprogrammeerd om een bepaald gedrag te vertonen Tussen deze beide extremen, de chemische samenstelling van de systemen en de input-output relaties aan de buitenkant van de systemen als *black box*, zijn nog een aantal mogelijkheden open De sterke AI moet specificeren voor welke mogelijkheid zij kiest. Dat is geen eenvoudige zaak. Van de computer zijn samenstelling, programma en input-output relaties bekend (tot op grote hoogte). Bij de mens is dat allemaal veel minder bekend Zoals reeds in 2.2.2 2 werd opgeworpen hoe onderscheid je de

specifieke eigenschappen van deze computer van de eigenschappen die elk intelligent systeem moet hebben, dus ook de mens?

In de loop van het bestaan van AI zijn verschillende oplossingen geboden voor het *grain*-probleem. Daarbij is een soort slingerbeweging te zien in de geschiedenis (18): aanvankelijk, in de vijftiger en begin zestiger jaren, koos men voor een oplossing met een zeer fijne korrel, vervolgens koos men een zeer grove korrel; en sinds een paar jaar is men weer teruggekeerd tot een oplossing met een middelfijne korrel.

2.3.1 Hersenen. Een fijnkorrelige vergelijking.

Het idee dat de mens een machine is is niet nieuw. Voor Descartes, en voor zijn volgelingen in het klooster van Port Royal, waren dieren gevoelloze automaten. Het menselijk lichaam was volgens hen net zo'n automaat, alleen was daar nog een menselijke, onstoffelijke geest aan toegevoegd. Zijn landgenoot La Mettrie (1709-1751) ging in zijn boek *L'homme machine* nog veel verder door nu ook de menselijke geest tot een functie van het mechaniek te maken en door de kwaliteit van zijn geestelijke vermogens te laten afhangen van de kwaliteit en de organisatie van het zenuwstelsel. De mens onderscheidt zich dan ook niet van het dier doordat hem een bijzondere, immateriële ziel is gegeven, maar door de meer verfijnde organisatie van zijn zenuwstelsel. La Mettrie verwerpt heel stellig de opvatting dat dieren gevoelloze automaten zouden zijn; ook de mens is in die zin geen automaat, maar zijn gevoeligheid en zijn redelijke vermogens kunnen volgens hem in de materie tot ontwikkeling komen (Thijssen 1982, 246).

Dit materialisme en mechanicisme van La Mettrie vond aanvankelijk weinig weerklank, maar werd gaandeweg meer en meer geaccepteerd. De nadruk lag bij hem op de organisatie van het zenuwstelsel. Deze aandacht voor het zenuwstelsel, en met name voor de hersenen, was ook in zijn tijd niet nieuw. Dualisten zoals Descartes, maar ook middeleeuwse artsen en wijsgeren voór hem, zochten in de hersenen het aangrijpingspunt of de zetel van de ziel. Maar voor materialisten was het al helemaal belangrijk aan te kunnen tonen dat de hersenen, waarvan de innige samenhang met mentale toestanden en gebeurtenissen bekend was, zelf in staat waren om de gevoeligheid en de redelijke

vermogens tot stand te brengen. Toen dan ook aan het einde van de veertiger jaren van deze eeuw computers gemaakt begonnen te worden, volledig materiële mechanismen die intelligentievereisende taken konden verrichten, leek dat een godsgeschenk voor de materialisten. Computers hadden geen lichamen - geen ledematen en geen zintuigen - maar computers kon men hersenen noemen. Dat gebeurde dan ook in de pers werden de eerste computers *'giant brains'* genoemd en er verschenen boeken met titels als *The Brain as a Computer* (George 1962) en *The Machinery of the Brain* (Wooldridge 1963) Wooldridge, ofschoon voorzichtig over de overeenkomst tussen neuronen en flip-flops, zegt:

> "... computers get their amazing results by the performance of a very large number of very simple processing steps. *This would also appear to be a valid description of the essence of* brain function " (Wooldridge 1963, 236).

En Fink schrijft.

> "The fact that such mental activities as reading, translating, problem solving and playing games can be reduced to numerical manipulation comes as a surprise to almost everyone not familiar with computer technology. The explanation is simple. whether we are conscious of the fact or not, the brain employs the rules of logic. In a very real sense, the brain *is* a computer and its computer-like functions can be imitated by machinery" (Fink 1966, 5).

Een mens is dan wel niet strikt gelijk aan een computer, maar wel aan een robot met een computer in zijn hoofd.

Die overeenstemming tussen hersenen en computer leek aanvankelijk heel ver te gaan. Beide zijn materiele mechanismen die toch met redelijke vermogens begaafd lijken te zijn. De eerste computers deden nog niet zulke geweldige dingen, maar de verwachtingen voor de toekomst waren hoog gespannen (19). Het computermodel voor de hersenen was evenwel ook gecorreleerd met werk in de neurofysiologie,

waar men gevonden had dat neuronen op een alles-of-niets wijze uitbarstingen van electriciteit voortbrengen. Een zo'n uitbarsting, *spike* genoemd, werd gezien als de basiseenheid van informatie in de hersenen die correspondeerde met de informatie-bit in de digitale computer. De gelijkheid tussen hersenen en computer werd met een zeer fijne 'korrel' gezien: weliswaar was hun chemische samenstelling verschillend, maar beide mechanismen waren opgebouwd uit digitale basiseenheden, de hersenen uit de alles-of-niets vurende neuronen, de computer uit digitale flip-flops Combinatie van deze informatiedragende basiseenheden leidde tot het vertonen van redelijke vermogens

Het enthousiasme van de eerste AI-werkers - de term bestond nog niet, ze werden cybernetici genoemd - was onmiddellijk zeer groot. Men legde zich erop toe neurale netwerken na te bouwen, minimale, zichzelf organiserende systemen. Men bouwde grote verzamelingen van gelijke basiscomponenten die zich, aanvankelijk zonder interne structuur maar geplaatst in een gunstige omgeving, uiteindelijk adaptief zouden gaan gedragen. Uit de evolutie van zo'n systeem zouden de redelijke vermogens ontstaan. Zulke systemen werden wel *perceptrons* genoemd. Het perceptrononderzoek ging uit van een extreem empiristische opvatting van de hersenen. De hersenen werden gezien als een *tabula rasa,* een bij de geboorte ongestructureerde verzameling neuronen, waaruit pas onder invloed van de omgeving organisatie en redelijke vermogens konden ontstaan. Al gauw waren al wel honderd research-groepen in Amerika bezig met deze perceptrons, leermachines of zelf-organiserende netwerken. Nog in 1971 werd geschreven dat deze groepen aangetoond hadden dat machines als deze konden leren, en wel sneller dan menselijke studenten (Toffler 1971, 186).

De kritiek op deze fijnkorrelige gelijkstelling van hersenen en computer was echter al veel eerder begonnen. Al in zijn beroemde artikel 'Computing Machinery and Intelligence' van 1950 zegt Alan Turing dat in het zenuwstelsel chemische verschijnselen minstens zo belangrijk zijn als elektrische Overeenkomsten tussen hersenen en computer zijn eerder te vinden in mathematisch analoge functies van beide dan in een overeenkomst tussen neuronen en flip-flops. En John von Neumann, de beroemde computerpionier naar wie de meest gebruikte soort digitale computers genoemd is, zegt in 1956 over de

gelijkheid van hersenen en computer dat de hersenen andere procedures gebruiken dan een computer.

> "... some of the steps being neural, that is digital, and others humoral, that is analog .."(Von Neumann 1956, 2077).

Allerlei niet-digitale factoren spelen in de hersenen een informatiedragende rol, zoals de diameter van de axonen (de verbindingen tussen neuronen), de frequentie van het vuren, de chemische samenstelling van de neurotransmitters. Het bleek gewoonweg onjuist de hersenen te beschouwen als een ongestructureerde verzameling van digitale flip-flops (neuronen). Rosenblith sluit in 1966 min of meer een periode af met de woorden:

> "We no longer hold the earlier widespread belief that the so-called all-or-none law from nerve impulses makes it legitimate to think of relays as adequate models for neurons . . Detailed comparisons of the organization of computer systems and brains would prove equally frustrating and inconclusive" (Rosenblith 1966, 247).

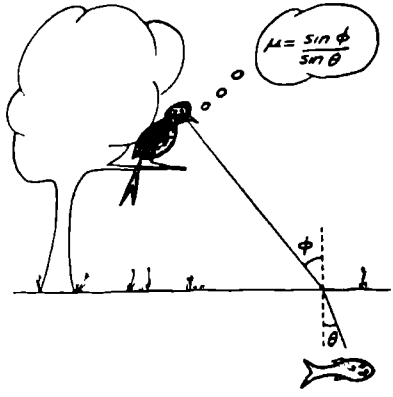
Minsky en Papert (1969, 4) zeggen dat de perceptronexperimenten over het algemeen teleurstellend waren en de meeste geschriften erover zonder wetenschappelijke waarde. De periode van fijnkorrelige gelijkstelling van mensen en computers op het niveau van de kleinste informatiedragende basiseenheden van hersenen en computers was afgesloten.

2.3.2. De Turing-test. Een grofkorrelige vergelijking.

Nu de gedetailleerde vergelijking van hersenen en computers niet langer veelbelovend leek, ging men over tot een zeer grofkorrelige gelijkstelling: men keek alleen nog naar de input-output equivalentie van mens en machine. Men probeerde in de AI eenvoudigweg intelligente machines te bouwen. In sommige onderzoeken werd

geprobeerd het gedrag van de machine zoveel mogelijk te laten overeenkomen met het gedrag van proefpersonen, en tot dat gedrag behoorden dan ook de hardop-denk protocollen van de proefpersonen. Maar over het geheel genomen was men erop uit om intelligente machines te maken zonder enige poging om het systeem eenvoudig, psychologisch of mens-achtig te maken Men had gezien dat de zelf-organiserende systemen weinig belovend waren, en trachtte werkende intelligente systemen te maken, desnoods gebaseerd op *ad hoc* mechanismen In deze benadering probeerde men de aard van intelligentie te begrijpen door intelligentie te creeren.

De gelijkstelling van mens en computer gebeurde nu uitsluitend op grond van input-output equivalentie. De banden met de neurofysiologie, die de AI aanvankelijk had, werden losgesneden. De geprogrammeerde computer vormde een existentiebewijs voor de mogelijkheid dat een volledig materieel mechanisme redelijke vermogens bezat, of in ieder geval intelligent gedrag leek te vertonen. Dat de binnenkant van een computer er heel anders uitzag dan de binnenkant van een mens deed er niet zoveel toe. Het werd ook niet problematisch geacht als de computer mogelijk op een andere manier tot zijn gedrag kwam dan de mens. Overigens wist men uberhaupt niet veel van de manieren waarop een mens tot zijn gedrag komt.



Figuur 2 (overgenomen uit Boden 1983, 12)

De hardop-denk protocollen van het probleemoplossen leverden niet zoveel op het overgrote deel van de denkactiviteiten tijdens het probleemoplossen bleek niet introspecteerbaar te zijn. Men nam om aprioristische redenen aan dat mensen en dieren op de een of andere manier computaties uitvoeren bij het produceren van intelligent gedrag, net als computers. Hoe kan dat gedrag anders tot stand komen (zie figuur 2)? Maar of die computaties voor het overige lijken op de computaties van de geprogrammeerde computer, daar meende men niets over te kunnen zeggen. De echte gelijkstelling van mens en computer bleef beperkt tot input-output equivalentie.

Nu was die input-output equivalentie niet onproblematisch. Wanneer zijn twee gedragingen gelijk? Het probleem dat we in 2.3 signaleerden dreigt zich te herhalen: men kan altijd vragen. "Gelijk in welk opzicht?" Tegenstanders van AI zijn geneigd over geprogrammeerde computers te zeggen: "Maar dat is niet echt rekenen", "Dat is niet echt converseren", "Dat is niet echt denken" enz. Men spreekt hier wel van de 'n+1-test' voor computerintelligentie. Wanneer de computer slaagt wordt de test verzwaard Volgens sommigen gebeurt dit omdat bij een begrip als 'denken' verklaren gelijk staat met 'wegverklaren'. Zo zegt Minsky hierover:

> "To me, "intelligence" seems to denote little more than the complex of performances which we happen to respect, but do not understand. So it is, usually, with the question of "depth" in mathematics. Once the proof of a theorem is understood, its content seems to become trivial ... we should not ... conclude that programmed computers therefore cannot think. For it may be so with *man,* as with *machine,* that, when we understand finally the structure and program, the feeling of mystery (and self-approbation) will weaken" (Minsky 1963, 447).

Alan Turing had in 1950 dergelijke problemen al voorzien. Volgens hem moet een antwoord op de vraag "Kunnen machines denken?" beginnen met definities van de termen 'machine' en 'denken' Turing ziet praktische problemen oprijzen bij het opstellen van zulke definities. Maar hij gelooft ook dat de vraag betekenisloos is, en dat bovendien

het gebruik van de termen tegen het eind van deze eeuw veranderd zal zijn. Turing probeert al deze problemen kort te sluiten:

> "Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words" (Turing 1950, 4).

De nieuwe vorm van de vraag kan beschreven worden in termen van een spel: het 'imitatiespel'. Dit wordt gespeeld door drie personen: een man (A), een vrouw (B), en een ondervrager (C). De ondervrager kan A en B niet zien of horen. Zij (of hij) moet door schriftelijke vragen erachter zien te komen wie de man en wie de vrouw is. Zij kent ze alleen bij de letters X en Y. Aan het eind moet ze zeggen "X is A (de man)" of "Y is A (de man)". A moet proberen C in de war te brengen, zodat ze de verkeerde identificatie maakt, terwijl B C moet helpen. B kan dus aan C schrijven: "Ik ben de vrouw, luister niet naar hem", maar dat helpt niet veel want A, die mag liegen, kan precies hetzelfde schrijven. Men kan zich nu afvragen: "Wat zal er gebeuren als een machine de rol van A neemt in dit spel? Zal de ondervrager dan even vaak fouten maken met de machine in het spel als wanneer het spel met een man en een vrouw gespeeld wordt?" Deze vraag vervangt dan de oorspronkelijke: "Kunnen machines denken?"

Dit imitatiespel staat nu algemeen bekend als de Turingtest, en wordt vaak vereenvoudigd tot de vraag: "Kan een ondervrager met schriftelijke vragen een onderscheid vinden tussen een mens en een computer?" En hoewel Turing zijn test voorstelde om discussies kort te sluiten, is er de afgelopen 35 jaar druk gediscussieerd over de Turingtest.

Het is moeilijk de kracht van de Turingtest te overschatten. Turing laat zien dat zijn spel bijna ieder veld van menselijke mogelijkheden kan omvatten. Een voorbeeld:

> Q: Please write me a sonnet on the subject of the Forth Bridge.
> A: Count me out on this one. I never could write poetry.
> Q: Add 34957 to 70764.

A: (Pause about 30 seconds and then gives an answer) 105621.

Q. Do you play chess?

A: Yes.

Q: I have K at my K1, and no other pieces. You have only K at k6 and R at R1 It is your move. What do you play?

A: (After a pause of 15 seconds) R-R8 mate (Turing 1950, 6).

Geen enkel programma is ooit voor de echte Turingtest omdat geen enkel programma zoveel mogelijke terreinen kan omvatten. Sommige programma's zijn op het eerste gezicht bijzonder indrukwekkend. Een steeds opnieuw gebruikt voorbeeld is het programma ELIZA van Weizenbaum, dat de rol van een non-directieve therapeut speelt:

P: "Men are all alike."

E: "In what way?"

P: "They're always bugging us about something or other "

E: "Can you think of a specific example?"

P: "Well, my boyfriend made me come here."

E: "Your boyfriend made you come here."

P: "He says I'm depressed much of the time."

E: "I am sorry to hear you are depressed."

P: "It's true, I am unhappy."

E: "Do you think coming here will help you not to be unhappy?"

P: "I need some help, that much seems certain."

E: "What would it mean to you if you got some help?"

P: "Perhaps I could learn to get along with my mother."

E: "Tell me more about your family." (Weizenbaum 1963).

Het aardige is dat ELIZA gemaakt is om te laten zien dat je met slimme trucs een heel plausibel programma kunt maken - al in 1963! Maar als je het programma gaat testen valt het meteen door de mand.

Het dichtst in de buurt van het passeren van de Turingtest kwam het programma PARRY van Kenneth Colby, dat een paranoïde patiënt

simuleert Psychiaters konden geen onderscheid maken tussen PARRY en een echte paranoicus, *ook al wisten ze dat een van beide een computer was* (Colby 1981) Toch was dat niet zo gek paranoici geven vaak stereotype antwoorden en de psychiaters waren voorzichtig omdat ze ook wisten dat een van beide een echte patient was Van verschillende zijden is dan ook beweerd dat PARRY niet geslaagd was voor een echte Turingtest, het duidelijkst door Joseph Weizenbaum Die zei dat hij nog een veel beter programma had om de Turingtest te winnen, en nog veel goedkoper ook Je had er zelfs helemaal geen computer voor nodig Het was een simulatie van een autistisch kind - je typt je vragen in en je krijgt geen antwoord: niet te onderscheiden van een echt autistisch kind (20)

Toch kent de Turingtest wel bezwaren Ten eerste kun je de test ook zien als een intelligentietest voor de ondervrager C. Wanneer er geen verschil wordt gevonden tussen A en B valt er niet te onderscheiden tussen de excellentheid van het computerprogramma en de ongevoeligheid van het meetinstrument (C) Het slagen voor de Turingtest mag dus niet gelden als *definitie* van 'kunnen denken'. Dat zou wat al te operationalistisch zijn En ten tweede is de Turingtest heel uitdrukkelijk een test voor input-output equivalentie Een scepticus zou zeggen "Maar input en output is ook alles wat we van onze medemensen zien - hoe weet je zeker of zij kunnen denken?" En voor een behaviorist zou, als antwoord op dit sceptische probleem, input-output equivalentie voldoende bewijs zijn voor het kunnen denken van machines (21) Maar enkel het matchen van input-output relaties kan niet het doel van wetenschap zijn· men zou ook wel willen weten waarom die relaties zo zijn Dat is de reden waarom men op den duur het behaviorisme in de psychologie, dat immers alleen naar input-output relaties kijkt, en alles wat daartussenin gebeurt beschouwt als een *black box,* onvruchtbaar vond En daarom ook vond men de gelijkstelling van mens en computer op het niveau van input-output equivalentie toch op den duur te grofkorrelig (zie ook Pylyshyn 1984, 121-122)

Vele AI-werkers zagen hun computerprogramma's toch als psychologische theorieen Zij wilden psychologische realiteit toekennen aan tenminste sommige van de processen die in hun programma's beschreven zijn, dat wil zeggen, ze wilden beweren dat sommige van die processen ook echt in de mens plaatsvinden. Volgens hen zijn mens en computer gelijk in meer dan alleen uiterlijk gedrag

Nu is zo'n gelijkstelling problematisch omdat, zoals boven gezegd, maar heel weinig interne processen voor mensen introspecteerbaar zijn. In de AI is het dan ook gebruikelijk om aan te nemen dat bijna alle computationele processen bij de mens onbewust zijn (zie b.v Dennett 1978b) Maar als dat zo is, dan is een gelijkstelling van mens en computer op het niveau van computationele processen volstrekt oncontroleerbaar. Wanneer twee verschillend geprogrammeerde computers input-output equivalent zijn met elkaar en met de mens, welke computationele processen hebben dan psychologische realiteit? Welke processen zijn bij de mens echt aanwezig (zij het onbewust), en welke zijn niet aanwezig maar leveren wel equivalent gedrag op?

De cognitiewetenschapper Zenon Pylyshyn heeft geprobeerd enige methodologische richtlijnen te geven voor dit probleem (Pylyshyn 1980, 1984) Uitgangspunt van deze richtlijnen is dat we een deel van het menselijk functioneren, zoals waarnemen, denken, handelen, het beste kunnen beschrijven in symbolische termen Dat betekent dat het gedrag van mensen beschreven kan worden als gestuurd door een *programma,* door regels die van toepassing zijn op symbolische representaties. Waarom pak ik mijn paraplu? Omdat ik geloof dat het regent, en niet nat wil worden, en denk dat een paraplu mij droog kan houden. Dit hele proces, dat uitmondt in mijn gedrag, kan beschreven worden als een reeks berekeningen, computaties, waarin een aantal symbolische representaties (van de regen, van mijn vooruitzicht om nat te worden) de argumenten vormen. Een ander deel van het menselijk functioneren kan niet zo beschreven worden. Waarom niesde ik zoeven? Omdat stofdeeltjes mijn neus binnendrongen die irritaties veroorzaakten, en dat leidde tot mijn niesbui Dit proces kan in puur fysische termen beschreven worden; er komt geen symbool in voor, de stofdeeltjes representeren niets. Dit proces is niet computationeel.

Pylyshyn wil de computationele modellen voor gedrag letterlijk nemen: het model werkt volgens het programma dat de mens ook echt gebruikt; dat wil zeggen, de in het programma beschreven processen hebben psychologische realiteit, ze bestaan echt in de mens en zijn niet slechts instrumenten om tot equivalente input-output relaties te komen. Maar om die psychologische realiteit te kunnen claimen is input-output equivalentie tussen model en mens niet genoeg. Het is echter mogelijk om empirisch onderscheid te maken tussen processen met dezelfde input-output functie.

Een cognitief model hoeft natuurlijk geen waarden te geven voor bijvoorbeeld bloeddruk, huidweerstand (GSR), of lichaamstemperatuur. De waarden van deze parameters zijn niet symbolisch, ze representeren niets. Het zijn eenvoudig intrinsiek fysische grootheden, en als zodanig horen ze niet thuis in de computationele processen. Maar zulke maten kunnen wel gezien worden als een aanwijzing voor bepaalde globale eigenschappen van de computationele processen, bijvoorbeeld van mentale belasting. Zo hangt de huidweerstand samen met de zweetafscheiding, en is het louter fysisch gevolg van een louter fysisch, niet computationeel proces. Maar bij een moeilijke of emotioneel beladen denktaak, op zich wèl een computationeel proces, kan als gevolg van die moeilijkheid of emotionele beladenheid, de zweetafscheiding toenemen, en daarmee de GSR veranderen. Ook in het dagelijks leven zeggen we dat iemand "het zweet uitbreekt" wanneer hij het in figuurlijke zin benauwd heeft. Zo ook kan de reactietijd bij een bepaalde taak niet gezien worden als een berekende respons, zoals het uitvoeren van de taak zelf dat wel is (22). Reactietijd kan gezien worden als gecorreleerd met een bepaalde globale eigenschap van het computationele proces, als index van de computationele complexiteit.

De interne, mentale of cognitieve, processen van mensen zijn niet waarneembaar, noch van buitenaf, noch (of slechts zeer zelden) door introspectie. Maar dat betekent niet dat tussen de verschillende input-output equivalente modellen niet onderscheiden kan worden. Er zijn bij mensen fysiologische en tijdsmaten te meten als index voor computationele complexiteit. Bij de computationele modellen is bekend welke processen plaatsvinden, zodat ook daar een maat voor complexiteit genomen kan worden. Wanneer nu mens en model voor iedere input dezelfde output produceren èn dezelfde maat voor

computationele complexiteit hebben, dan noemt Pylyshyn ze complexiteitsequivalent De complexiteitsequivalentie relatie is een verfijning van de zwakke input-output equivalentie. Het is een stap in de richting van *sterke equivalentie,* dat wil zeggen die equivalentie waarbij men mag claimen dat mens en model werkelijk dezelfde processen gebruiken Criteria voor complexiteitsequivalentie helpen ons te kiezen tussen verschillende input-ouput equivalente modellen. Andere criteria zijn die welke aanwijzingen geven over tussenuitkomsten en duidelijke subcomponenten van het proces. Wanneer bijvoorbeeld een rekenmachine bij het uitrekenen van 4 x 18 als tussenuitkomsten heeft 36 en 54, terwijl wij mensen als tussenresultaat 32 en 40 hebben, dan bereiken we de uitkomst, 72, niet op dezelfde manier, en zijn onze rekenprocessen niet sterk equivalent. Zo kan men steeds verder komen in de richting van sterke equivalentie tussen mens en machine.

Het is belangrijk in te zien dat in de voorafgaande overwegingen over equivalentie onderscheid is gemaakt tussen geobserveerde gegevens die direct zijn toe te schrijven aan het computationele proces, zoals uitkomsten en tussenuitkomsten, en gegevens die zijn toe te schrijven aan de fysische eigenschappen van het systeem dat het computationele proces uitvoert, zoals GSR, temperatuur, reactietijd. Deze laatste maten zijn, zoals gezegd, niet symbolisch en niet computationeel, maar ze zeggen wel wat *over* die symbolische computaties. Pylyshyn zegt hierover:

> "The distinction between behavior being governed by symbolic representations and behavior being merely exhibited by a device in virtue of the causal structure of that device is one of the most fundamental distinctions in cognitive science" (Pylyshyn 1980, 120).

Het is namelijk niet langer vol te houden, zoals men in de periode van de grofkorrelige vergelijking wel meende, dat de fysische eigenschappen van het systeem dat het programma uitvoert helemaal niet van belang zouden zijn. We hebben al gezien dat bepaalde fysische grootheden iets kunnen zeggen *over* globale eigenschappen van de computationele processen. Maar bepaalde fysische eigenschappen van

het onderliggend systeem kunnen ook eigenschappen van de computationele processen *bepalen*. Het gaat daarbij niet om de fysische of chemische eigenschappen op een zeer fijnkorrelig niveau, maar om bepaalde functioneel gespecificeerde eigenschappen op het niveau van de structuur van het systeem. Pylyshyn spreekt van de eigenschappen van de *functionele architectuur*. Hij noemt een voorbeeld: een bepaald AI waarnemingsprogramma vertoonde bij de werking een effect dat heel veel leek op de Muller-Lyer illusie het programma beoordeelde een lijn met pijlen aan de uiteinden, zoals ⟷ korter dan een lijn met vorken aan de uiteinden, zoals ≻——≺ .
Het programma zou een verklaring kunnen geven voor deze illusie Nu was dit verschijnsel het gevolg van het feit dat het fysische systeem dat het programma uitvoerde een diameter-beperkte 'zoeker' gebruikte om lijnen te herkennen. Deze zoeker zocht de lijnen af naar aanwijzingen voor bepaalde soorten hoeken aan de uiteinden, zoals pijlen, vorken, L-vormige hoeken enz. De aanwijzingen voor een pijl zijn eerder aanwezig voor de zoeker dan die voor een vork, omdat de secundaire pijllijnen al in het zoekerveld verschijnen voor de punt van de pijl Zo herkent het systeem het eind van een lijnstuk eerder wanneer het met een pijl eindigt dan wanneer het een lijnstuk met een vork afzoekt. Daarom geeft het systeem in het eerste geval een kortere lengteschatting. Het is in dit geval dus niet de aard van het gebruikte programma die de illusie verklaart, maar een eigenschap van het hebben van een zoeker met beperkte diameter. Of de mens ook zo'n zoeker heeft in zijn functionele architectuur moet onafhankelijk getoetst worden.

Het is zo dat ieder computationeel model, wil het letterlijk genomen worden, vooronderstellingen moet maken over het onderliggende mechanisme, over de eigenschappen van dat mechanisme die zelf niet geprogrammeerd zijn. De aard van een visuele 'zoeker' is er een van. De capaciteit van het geheugen is een andere eigenschap van de functionele architectuur die verstrekkende gevolgen heeft voor computationele processen. Een systeem met een zeer beperkt korte-duur-geheugen bijvoorbeeld, gebruikt voor allerlei taken heel andere strategieen en maakt heel anderssoortige fouten dan een systeem met een veel grotere capaciteit Een beperkt korte-duur-geheugen dwingt het systeem zijn strategieen te baseren op heel bepaalde basisfuncties.

Dit valt goed te illustreren aan de hand van het verschil tussen 'hoofdrekenen' en 'cijferen' Bij hoofdrekenen gaat het om het uitrekenen van sommen 'uit het hoofd', bekende strategieen daarbij zijn tijdelijk afronden of het opsplitsen van getallen in termen van guldens en kwartjes: 100-tallen, 25-tallen, 12,5-tallen enz. 'Cijferen' is rekenen op papier, hierbij spelen geheugenbeperkingen geen rol. Het papier fungeert immers als extern geheugen. De normale strategie hierbij is het rekenen in kolommen van rechts naar links, waarbij alleen het transport moet worden onthouden. De beperktheid van het korte-duur-geheugen dwingt ons om de informatie waar we mee moeten werken in bepaalde handzame 'brokken', zogenaamde *chunks,* op te delen De manier waarop we de informatie in *chunks* opdelen is afhankelijk van denkprocessen, en verklaarbaar in termen van regels en representaties Maar het opslaan en terughalen van *chunks* uit het geheugen zelf is een primitieve operatie, een basisoperatie waar we mee moeten werken. En het feit dat die *chunks* niet te groot mogen zijn, ongeveer zeven elementen, is een primitieve eigenschap van die primitieve operatie, een eigenschap van onze functionele architectuur Meer in het algemeen, de functionele architectuur van een bepaald systeem bepaalt welke functies in een programma, in de computationele processen, *primitief* zijn. Bij de mens bepaalt de functionele architectuur de basisoperaties van de mentale processen, die zelf geen procesverklaring meer krijgen. Die basisoperaties of primitieve functies zijn zelf biologisch verklaarbaar, en niet in termen van regels en representaties. Bijvoorbeeld, zo'n heel bepaalde geheugen-basisoperatie is het directe gevolg van de beperktheid van het korte-duur-geheugen, en die beperktheid is een biologisch verklaarbare eigenschap van de hersenen.

Aangezien we dus bij het programmeren rekening moeten houden met de aard van het systeem dat het programma uitvoert bestaat het maken van computermodellen in de psychologie eigenlijk uit twee fasen eerst moet de functionele architectuur van de mens in de machine nagemaakt worden, en dan moet op die machine het gepostuleerde computationele proces uitgevoerd worden Dat namaken van de functionele architectuur in een machine kan op twee manieren gebeuren. Men kan de *hardware* van de machine zo bouwen dat de primitieve functies er direct uit volgen. De functionele architectuur *is* dan de manier waarop

de machine gebouwd is. Maar men kan de machine ook *programmeren alsof* de primitieve functies direct eruit volgen. Iedere programmeertaal van hoog niveau kent primitieve functies. Met behulp van deze primitieve functies moeten de procedures van het programma samengesteld en gedefinieerd worden, het zijn de basisblokken van het programma Het lijkt dan of de machine waarop men werkt zo gebouwd is dat die primitieve functies van de programmeertaal een gevolg zijn van de *hardware* (vgl. 2.2.1). Maar meestal zijn er verschillende lagen van programma-interpretatie tussen de programmeur en de *hardware* van de machine, die een heel andere architectuur kan hebben dan de primitieve functies van de programmeertaal zouden doen vermoeden. De functionele architectuur wordt dan nagemaakt of *geëmuleerd* door programma's op een lager niveau. Voor de programmeur maakt het evenwel niet uit of de functionele architectuur waar hij mee moet werken rechtstreeks in de bouw van de hardware zit, of op een ander soort machine is geëmuleerd.

Bij de mens wordt de functionele architectuur geacht rechtstreeks uit de biologische structuur te volgen. De primitieve functies van onze computationele processen moeten zelf biologisch verklaard kunnen worden. Ze moeten dus ook voor alle mensen universeel zijn, en (relatief) onveranderlijk. Hoe weet je nu of bij mensen bepaalde primitieve functies echt behoren tot de functionele architectuur, of dat ze, door gewoonte of opvoeding of wat ook, door programmeren op een lager niveau geëmuleerd zijn? Bijvoorbeeld, wij zien (tweedimensionele) tekeningen als afbeeldingen van driedimensionele voorwerpen. Kunnen we niet anders, omdat onze zintuigen op een bepaalde manier gebouwd zijn, of is dat een culturele verworvenheid? Zouden er culturen zijn waar men tekeningen uitsluitend ziet als tweedimensionele patronen? Bij computers waren beide mogelijkheden open, functionele architectuur als hardware en geëmuleerd. Waarom niet bij mensen?

Pylyshyn stelt een criterium voor om te bepalen wat bij mensen tot de functionele architectuur behoort, in de zin van biologisch verklaarbare architectuur: zo'n primitieve architecturale functie moet *cognitief ondoordringbaar* zijn. Daar wil hij mee zeggen dat zo'n functie ongevoelig moet zijn voor informatie die buiten die functie beschikbaar is Om een verhelderend voorbeeld van Fodor te gebruiken (Fodor 1983, 71): stel, je kent me al heel lang, en bent

overtuigd geraakt van de zachtaardigheid van mijn karakter. Meer in het bijzonder, je bent ervan overtuigd dat ik onder geen voorwaarde mijn vinger in je oog zal steken. Toch, als ik mijn vinger snel naar je oog toebeweeg, doe je dat oog dicht Geen informatie over mij en mijn zachtaardigheid kan daar iets aan veranderen. Deze reflex is cognitief ondoordringbaar.

Nu is het zoeken naar vaste architecturale functies niet gemakkelijk volgens Pylyshyn. Hij zegt bijvoorbeeld.

> "... the detection of everything from distance information to one's grandmother appears to be cognitively penetrable" (Pylyshyn 1980, 131).

Ook in de reacties op zijn artikel uit 1980 wordt opgemerkt dat vrijwel geen architecturale functies ondoordringbaar zijn Zelfs zuiver biologische processen zoals de hartslag en de spijsvertering zijn gevoelig voor informatie die buiten die processen beschikbaar is, en zijn door sommige mensen zelfs willekeurig te beïnvloeden en te veranderen. Bovendien zijn architecturale functies alleen voor zeer korte duur 'vast'. Op de wat langere duur blijkt dat de architectuur zelf, bijvoorbeeld het zenuwstelsel, voortdurend verandert onder invloed van informatie van buitenaf (zie b.v. P. Churchland 1980 en P.S. Churchland 1980) (23). Pylyshyn geeft toe dat het onderscheid tussen cognitief doordringbare en cognitief ondoordringbare functies in de praktijk moeilijk aan te brengen is, en ook dat beide soorten functies interacteren. Maar hij houdt vol dat het onderscheid principieel is. Alleen cognitief doordringbare functies zijn gevoelig voor het juiste soort invloed van buitenaf, voor invloed van informatie Een niet-cognitief en niet-computationeel proces als de spijsvertering is niet rechtstreeks gevoelig voor informatie, maar wel voor bepaalde aspecten van cognitieve processen die zelf gevoelig zijn voor informatie. Men kan veranderingen in de spijsvertering het beste zien als een maat voor mentale belasting, net als huidweerstand. Spijsvertering noch huidweerstand (GSR) zijn zelf cognitieve processen.

Overigens vindt Pylyshyn het niet verwonderlijk dat zo weinig functies cognitief ondoordringbaar blijken te zijn als men bedenkt hoe

enorm flexibel de mens is. Onze functionele architectuur legt ons slechts minimale beperkingen op in vergelijking met de inflexibiliteit van lagere diersoorten.

Pylyshyn's voorstel om de functionele architectuur te onderscheiden van de programma's die voor de mentale processen gebruikt worden levert een werkverdeling op voor de psychologie De computationele psychologie zoekt uit wat het mentale programma is, en verklaart menselijk handelen in termen van regels en representaties. De experimentele psychologie zoekt uit wat de functionele architectuur is op grond van universalia in het gedrag. En de neurowetenschappen verklaren fysiologisch en biologisch waarom de functionele architectuur zo is. In een machine kan de functionele architectuur nagebouwd of geemuleerd worden, en dan kan het gepostuleerde programma erop uitgevoerd worden. De mens is dan gelijk aan de machine opwaarts vanaf de functionele architectuur (de manier waarop de architectuur is samengesteld verschilt: biologische *wetware* of computer *hardware*).

Pylyshyn wil de mens-machine gelijkheid zien met middelfijne korrel. Alleen input-output equivalentie is hem te grofkorrelig; hij geeft methodologische richtlijnen om te kunnen komen tot sterke equivalentie, waarbij je kunt zeggen dat mensen en computers niet alleen input-output equivalent zijn maar ook dezelfde cognitieve processen gebruiken om tussen input en output te medièren. Bovendien zijn ook de basisbouwblokken van die cognitieve processen bij mens en juist geprogrammeerde machine gelijk; ze hebben dezelfde functionele architectuur. Hij zegt echter niet dat mens en machine zozeer gelijk zijn dat ook hun basisbouwblokken op fysiologisch niveau gelijk zouden zijn, op het niveau van neuronen en flip-flops. Zo'n vergelijking is hem weer te fijnkorrelig.

2.3.4. Pylyshyn versus Fodor over cognitieve ondoordringbaarheid.

In deze paragraaf wordt het criterium van cognitieve ondoordringbaarheid besproken als grens voor het domein van AI en cognitieve psychologie.

Pylyshyn's methodologische voorstellen geven aan met welke korrel van oplossing men moet kijken om de gelijkheid tussen mens en machine

te zien Daarmee kan het argument voor een fysicalistische oplossing van het lichaam-geest probleem, zoals die gangbaar is in de cognitieve psychologie, aanzienlijk verfijnd worden Het argument luidde, in zijn grofste vorm "Computers zijn net als mensen Het gedrag van computers kan in louter fysische termen beschreven worden, maar ook in intentionele termen. Zo ook zijn mensen volledig fysische systemen wier gedrag in fysische termen en in intentionele termen beschreven kan worden" Met Pylyshyn's methodologische voorstellen kan de eerste zin van dit mini-argument gepreciseerd worden tot. "Mensen hebben dezelfde functionele architectuur en cognitieve processen als (juist geprogrammeerde) computers, menselijke cognitie is in letterlijke zin computationeel".

Zijn voorstellen doen echter nog meer. Zijn criterium van cognitieve ondoordringbaarheid geeft niet alleen een grens aan van hoe diep je in een systeem (mens of machine) moet kijken om de gelijkheid te zien, maar ook een grens van het domein van cognitieve psychologie en AI. De cognitief doordringbare processen, waarvan de basisoperaties bepaald worden door de functionele architectuur, zijn gelijk bij mens en machine. Tevens zijn het die cognitief doordringbare processen die het domein vormen van de cognitieve psychologie en van de AI. Het verklaren van de functionele architectuur is werk voor de neurofysiologie, en het emuleren van die architectuur in een machine is werk voor ingenieurs

Tenslotte geeft Pylyshyn's nadruk op de functionele architectuur van een systeem aan dat men in de cognitieve psychologie van mening is dat de hersenen niet een ongestructureerde verzameling neuronen kunnen zijn, en dat een rijk repertoire van cognitieve processen alleen uitgevoerd kan worden door een duidelijk gestructureerd systeem Ten tijde van de perceptrons (zie 2.3 1) meende men nog dat een volstrekt ongestructureerd systeem ten grondslag lag aan onze cognitieve processen, maar in de AI kwam men al gauw tot de conclusie dat perceptrons niet erg veel konden leren, en dat de hersenen dus anders moeten werken. Geprogrammeerde machines met een duidelijke structuur bleken inderdaad veel meer te kunnen. Op een heel ander terrein had Chomsky laten zien dat het leren van een taal niet beschreven kan worden als het leerproces van een *tabula rasa* die de juiste stimuli krijgt aangeboden. Volgens Chomsky moet een kind al een

aangeboren structuur voor taal hebben, om in staat te zijn om in een paar jaar op grond van een zeer beperkt taalaanbod een volledig grammaticaal taalvermogen op te bouwen (zie b v. Chomsky 1980).

De filosoof Jerry Fodor heeft in zijn monografie *The modularity of mind* (1983) het begrip 'functionele architectuur' opgepakt en een theorie voorgesteld voor de structuur van de *mind*. Hij verwijst voor het gebruik van de *term* 'functionele architectuur' één keer naar Pylyshyn's artikel uit 1980 (Fodor 1983, 31); en Pylyshyn zegt in zijn boek uit 1984, dat gebaseerd is op dat artikel, in een voetnoot dat Fodor's monografie op een *minor point* misschien zijn eigen theorie tegenspreekt. Maar ik zie een belangrijk verschil tussen beider theorieen ondanks het gebruik van gelijke termen en van oppervlakkig bezien dezelfde begrippen Dat belangrijke verschil heeft te maken met het criterium van cognitieve ondoordringbaarheid - Fodor spreekt van informationele inkapseling - voor de afbakening van het domein van de cognitieve psychologie en de AI.

Fodor ontwikkelt een nieuwe versie van de *vermogens-psychologie.* Zo'n theorie stelt dat er vele fundamenteel verschillende soorten psychologische mechanismen (vermogens) gepostuleerd moeten worden teneinde de feiten van het mentale te kunnen verklaren. Men kan in de vermogens-psychologie twee varianten onderscheiden: een 'horizontale' en een 'verticale' vermogens-psychologie. De *horizontale* is de meest bekende. cognitieve processen vertonen de interactie van zulke vermogens als bijvoorbeeld het geheugen, het voorstellingsvermogen, de aandacht, het oordeelsvermogen, de waarneming. Deze vermogens zijn onafhankelijk van het onderwerp van de cognitieve processen: hetzelfde oordeelvermogen speelt een rol in, bijvoorbeeld, het analyseren van een filosofische stelling alsook in de perceptuele herkenning van een vioolconcert van Bach, of het innemen van een moreel standpunt inzake de kruisraketten. Hetzelfde geldt voor de andere vermogens: op ieder gebied worden dezelfde vermogens toegepast. De functionele architectuur van de geest heeft, volgens deze leer, een horizontale indeling.

De *verticale* versie van de vermogens-psychologie gaat hier tegenin. Volgens deze leer is er niet eén soort intelligentie, één soort oordeelsvermogen, één soort geheugen enz. Fodor citeert hier Franz Joseph Gall (1758-1828), de grondlegger van de verticale vermogens-

> "Perception and memory are only attributes common to the fundamental psychological qualities, but not faculties in themselves, and consequently they can have no proper centres in the brain" (geciteerd in Fodor 1983, 16)

Gall beweert dat er voor ieder domein van cognitieve processen een apart psychologisch mechanisme bestaat, en dat die mechanismen geen vermogens (zoals het geheugen) delen Geheugen voor muziek is een eigenschap van de muzikale faculteit, en heeft niets te maken met geheugen voor literatuur, een eigenschap van de literaire faculteit Volgens Gall heeft de functionele architectuur van de geest een verticale structuur (24)

Fodor nu stelt een gemengd horizontale en verticale vermogens-psychologie voor volgens hem is de functionele architectuur van de geest opgebouwd uit zowel een aantal verticale als een aantal horizontale vermogens Hij onderscheidt in de architectuur een aantal input-systemen en een groot centraal systeem De input-systemen zijn volgens hem een soort verticale vermogens, Fodor noemt ze *modulair*. Dat wil zeggen dat ze onafhankelijk van elkaar en van het centrale systeem werken Ze zijn cognitief ondoordringbaar - hij noemt ze *informationeel ingekapseld*. Het centrale systeem is volgens Fodor niet modulair, en mogelijk gemedieerd door horizontale vermogens (hij laat zich niet uitvoerig uit over horizontale vermogens)

Fodor geeft een aantal argumenten en vele empirische aanwijzingen voor zijn theorie, en voor wat hij precies bedoelt met 'input-systemen' en met 'modulariteit' Input-systemen zijn er, het woord zegt het al, voor de input van stimuli in het systeem (organisme) Er zijn input-systemen voor de vijf zintuigen en voor de taal Dat taal bij de input-systemen wordt gerekend lijkt misschien wel wat vreemd Maar Fodor laat zien dat voor taal dezelfde eigenschappen gelden als voor de andere input-systemen Er zijn input-systemen voor de vijf zintuigen en voor de taal, maar hun aantal is veel groter dan zes Binnen, en mogelijk ook in dwarsdoorsnede van, de traditionele zintuigen zijn vele hooggespecialiseerde computationele mechanismen Hun taak is het om hypothesen te formuleren over de bronnen, in de buitenwereld, van de

binnenkomende stimulatie Input-systemen hebben een aantal eigenschappen gemeen op grond waarvan ze modulair zijn. ze zijn domeinspecifiek en informationeel ingekapseld, ze kunnen daardoor zeer snel werken maar geven daarom ook slechts een 'oppervlakkige' output waarvoor niet veel tijd en achtergrondinformatie vereist is, er is geen centrale toegang tot hun berekeningen en ze zijn geassocieerd met vaste neurale architectuur (ze zijn gelocaliseerd in de hersenen) en met specifieke storingsverschijnselen.

In tegenstelling tot de input-systemen zijn de centrale processen niet domein-specifiek en niet informationeel ingekapseld In de centrale processen wordt wat de verschillende input-systemen afleveren met elkaar in verbinding gebracht· wat we zien en horen, wat we aan talige informatie binnenkrijgen, wat we voelen, proeven en ruiken draagt allemaal bij tot hoe we geloven dat de wereld is Al die informatie moet met elkaar in verband worden gebracht: de processen die dat doen kunnen dus niet domein-specifiek en informationeel ingekapseld zijn. Fodor vergelijkt zijn centrale processen met de Aristotelische notie van een *sensorium commune:* een 'plaats' waar alle informatie van de domeinspecifieke sensoren samenkomt en vergeleken kan worden (zie ook Verwey 1984). De centrale processen zijn ook, omdat ze zoveel informatie in overweging moeten nemen, langzaam Bij het afwegen wat de beste hypothese is over hoe de wereld, of enig aspect van de wereld, is, kan alle mogelijke informatie relevant zijn. Fodor maakt een vergelijking met de wetenschap· de feiten die relevant zijn voor de bevestiging van een wetenschappelijke hypothese kunnen afkomstig zijn van *overal* in het veld van eerder vastgelegde empirische (of andere) waarheden. *Alles* wat de wetenschapper weet is, in principe, relevant om vast te stellen wat hij nog meer zou moeten aannemen. Dat geldt ook voor onze alledaagse, vaak onbewuste hypothesen over de wereld *alles* wat we weten is, in principe, relevant voor de beoordeling van nieuwe hypothesen en van nieuwe informatie, en voor de *updating* van onze kennis op grond daarvan. De centrale processen werken globaal, ze zijn niet domein-specifiek of informationeel ingekapseld.

Het interessante van Fodor's monografie in dit verband is het volgende· in zekere zin gebruikt Fodor hetzelfde criterium als Pylyshyn om verschillende interne processen te onderscheiden· Fodor's

'informationele inkapseling' is hetzelfde als Pylyshyn's 'cognitieve ondoordringbaarheid'. Maar voor het overige verschillen beiden flink van mening over die informationele inkapseling of cognitieve ondoordringbaarheid, en niet alleen op een *minor point*. De meningsverschillen betreffen twee punten ten eerste de vraag *hoeveel* processen informationeel ingekapseld of cognitief ondoordringbaar zijn, en ten tweede het gebruik van inkapseling of ondoordringbaarheid als criterium voor het afgrenzen van het domein van cognitieve psychologie en AI.

Volgens Pylyshyn zijn bijna alle processen cognitief doordringbaar, terwijl volgens Fodor alle input-systemen informationeel ingekapseld zijn Hoe is zo'n meningsverschil over een vrij feitelijk punt mogelijk? Pylyshyn's mening vindt waarschijnlijk meer aanhang in de cognitieve psychologie dan die van Fodor Niet alleen is de notie van de theoriegeladenheid van perceptie wijdverbreid, maar zowel de experimentele psychologie als de AI leveren vele illustraties van de effecten van informatiefeedback op input-operaties. verwachtingen bepalen deels wat we waarnemen. *Top-down* benaderingen verdienen vaak de voorkeur boven *bottom-up* benaderingen.

Tegenover Pylyshyn's mening stelt Fodor dat de input-systemen wel informationeel ingekapseld zijn, en voornamelijk *bottom-up* werken. Hij laat zien dat de uitkomsten van een aantal perceptie-experimenten anders uitgelegd kunnen worden - waar er sprake is van informatiefeedback is dat alleen een informatiestroom *binnen* het input-systeem - en verwijst in de AI naar de *bottom-up* benadering van het perceptie-onderzoek van Marr (b.v. Marr en Poggio 1977, Marr 1982). Maar Fodor heeft vooral a priori redenen voor het afwijzen van een volledige informatiefeedback naar de input-systemen, een volledige theoriegeladenheid van perceptie (zie ook Fodor 1984b). Perceptie moet ons informatie geven over hoe de wereld *is*, en die informatie moet *snel* komen. Teveel verwachting-gestuurde perceptie leidt tot *wishful seeing*, en zou al gauw dodelijk zijn We moeten ook onverwachte en onaangename dingen kunnen identificeren, en als de input-systemen onbeperkte toegang hebben tot alles wat we weten, zou dat identificeren wel eens te lang kunnen duren. De input-systemen leveren volgens Fodor snel een ondiepe input af; pas daarna, in het centrale systeem, vinden inferentieprocessen plaats waarbij alles wat

men weet betrokken kan worden De input-systemen zelf zijn informationeel ingekapseld

Fodor's opvatting over de inkapseling van input-systemen wijkt zoveel af van Pylyshyn's ideeen, dat het vreemd is dat Pylyshyn hier spreekt over een *minor point* en er verder het zwijgen toe doet. Maar nog veel belangrijker is het verschil tussen beider meningen over het gebruik van ondoordringbaarheid of inkapseling als criterium voor het afgrenzen van het domein voor de cognitieve psychologie en de AI. Voor Pylyshyn is cognitieve ondoordringbaarheid een criterium om te bepalen wat niet meer tot het domein van de cognitieve psychologie en de AI hoort Een cognitief ondoordringbaar proces is niet meer verklaarbaar in termen van regels en representaties, maar enkel biologisch verklaarbaar volgens hem Zo'n proces vormt een primitieve operatie in de computationele processen die de cognitieve psychologie moet beschrijven (en de AI moet nabouwen), en kan niet zelf nog als computationeel proces geanalyseerd worden.

Voor Fodor evenwel kunnen de processen van de input-systemen, die informationeel ingekapseld zijn (cognitief ondoordringbaar), zeer wel geanalyseerd worden als computationele processen. De processen die zich afspelen in de input-systemen zijn ons niet bewust - er is immers geen centrale toegang tot die processen - maar ze zijn wel computationeel Het oor voert Fourrier-analyses uit op de binnenkomende geluidsgolven, het oog betrekt de eigen bewegingen in de berekening van bewegingen in de buitenwereld (25).

Fodor wijst erop dat juist het beste werk in de AI gedaan is op het gebied van de input-systemen, van perceptie en taal. Cognitieve ondoordringbaarheid of informationele inkapseling vormt volgens hem helemaal geen criterium om uit te maken wat in het domein van de cognitieve psychologie en de AI valt.

De verschillen tussen Fodor en Pylyshyn over cognitieve ondoordringbaarheid zijn op zichelf interessant en in de literatuur (voor zover ik kan zien) nog niet besproken. Voor mijn betoog zijn die verschillen relevant voor zover het gaat over het domein van de AI. Pylyshyn meent dat alle cognitieve processen het domein vormen van de AI, en zijn criterium van cognitieve ondoordringbaarheid moet uitmaken welke processen niet meer cognitief zijn, en dus niet meer te analyseren zijn in computationele termen. Fodor laat niet alleen zien

dat Pylyshyn's demarcatiecriterium niet geschikt is, hij gaat veel verder. Hij laat niet alleen zien dat de cognitief ondoordringbare processen *ook* geschikt zijn voor analyse in computationele termen, hij oppert zelfs de mogelijkheid dat *alleen* cognitief ondoordringbare processen geschikt zijn voor zo'n analyse, en dat de doordringbare cognitieve processen juist niet het domein van de AI vormen. Ik bespreek zijn argumenten voor die mogelijkheid in de volgende paragraaf.

2.4. De mens-machine gelijkheid. Het frame-probleem

In 2.2.1 hebben we gezien dat we een computer kunnen bouwen voor elk proces dat men een effectieve procedure kan noemen. En een effectieve procedure is een verzameling regels die de machine van moment tot moment exact aangeven wat te doen. Nu kan men zich natuurlijk afvragen of alle cognitieve processen effectieve procedures genoemd kunnen worden. Is het mogelijk om al onze cognitieve processen te formaliseren tot een verzameling exacte regels? De sterke AI moet ervan uitgaan dat dit in principe mogelijk is. Ook Pylyshyn gaat ervan uit dat in principe alle cognitieve processen het domein van de AI vormen.

Alleen als computers in principe alles kunnen wat mensen kunnen (op cognitief gebied) kan het bestaan van computers een rol spelen in een argumentatie voor de fysicalistische oplossing voor het lichaam-geest probleem die gangbaar is in de cognitieve psychologie. Wanneer men Pylyshyn's werk in aanmerking neemt, luidt die argumentatie: "Mensen hebben dezelfde functionele architectuur en dezelfde cognitieve processen als juist gebouwde en geprogrammeerde computers. Het gedrag van computers kan in louter fysische termen beschreven worden, maar ook in intentionele termen. Zo ook zijn mensen volledig fysische systemen, wier gedrag in fysische termen en in intentionele termen beschreven kan worden". Als het echter zo is dat belangrijke cognitieve processen niet programmeerbaar zijn, niet in het domein van de AI vallen, dan kan het bestaan van computers niet meer een rol spelen in de argumentatie. Mensen zijn dan, in een zeer relevant opzicht, *niet* net als computers; en dan behoeft het feit dat menselijk

gedrag op twee heel verschillende manieren beschreven en verklaard kan worden nog steeds een verklaring. Natuurlijk kan men nog wel volhouden dat mensen volledig fysische systemen zijn; maar het bestaan van computers geeft aan die stelling dan geen empirische steun

Fodor oppert de mogelijkheid dat de centrale, informationeel niet ingekapselde denkprocessen zich niet lenen voor analyse in termen van effectieve procedures, van formele regels. De reden daarvoor is *precies dat ze informationeel niet ingekapseld zijn.* Voor de centrale denkprocessen kan *alle* informatie relevant zijn: ze zijn globaal. Er lijkt geen manier te zijn om het soort informatie te beperken dat invloed heeft op, of beïnvloed wordt door, centrale probleemoplossingsprocessen. Een voorbeeld van dit soort moeilijkheden is wat in de AI bekend staat als het *frame*-probleem, het probleem om een 'frame', een raamwerk, te plaatsen rond de verzameling kennis die bijgewerkt moet worden in het licht van bepaalde nieuwe informatie De nieuwe informatie eenvoudig bij de aanwezige kennis optellen is niet juist: sommige oude informatie is ongeldig geworden Alle aanwezige kennis nazien om te kijken wat er nog geldig is en wat niet meer is ondoenlijk Maar het is ook onmogelijk om van te voren vast te leggen welke informatie er herzien zal moeten worden. En het heeft geen zin om te zeggen dat alle relevante kennis herzien moet worden, omdat letterlijk alle kennis relevant kan zijn voor de centrale processen

Fodor laat hiermee zien dat de centrale, globale denkprocessen zich juist niet lenen voor een formele analyse. Hij noemt dit *"Fodor's First Law of the Nonexistence of Cognitive Science"* · hoe globaler een cognitief proces is, des te minder men het begrijpt. Heel erg globale processen, zoals redeneringen naar analogie, worden helemaal niet begrepen (Fodor 1983, 107) Maar de locale, informationeel ingekapselde processen van de input-systemen lenen zich heel goed voor een formele analyse in termen van regels en representaties Ook expertsystemen zijn voorbeelden van succesvolle formele analyse en dus van succesvolle programmering. De redeneerprocessen van expertsystemen zijn niet processen van input-systemen. Toch weerlegt dat Fodor's stelling niet. Expertsystemen zijn bij uitstek programma's van locale, domeinspecifieke, informationeel ingekapselde processen Bij de expertsystemen speelt het frame-probleem juist geen rol: ze zijn *per*

definitie slechts werkzaam op een nauwkeurig afgegrensd domein. Geen pogingen worden zelfs maar ondernomen om de domeinen van de verschillende expertsystemen uit te breiden of te combineren. Expertsystemen vormen een technologische triomf voor de AI, maar ze laten niet zien dat normale centrale denkprocessen formaliseerbaar zijn.

Fodor is niet de eerste die zich afvraagt of alle cognitieve processen wel formaliseerbaar zijn. Al twintig jaar geleden begon de filosoof Hubert Dreyfus zijn niet aflatende aanvallen op de AI (1965, 1972, 1979). Hij wees er voortdurend op dat de globale denkprocessen niet formaliseerbaar zijn, dat relevantie niet programmeerbaar is, dat kennis van de wereld niet analyseerbaar is in termen van context-vrije, atomische gegevens, maar dat alle kennis met elkaar samenhangt. Er zijn, zoals de ondertitel van één van zijn boeken ook zegt, grenzen aan de AI. Dreyfus' argumentatie gaat ietwat anders dan die van Fodor. Hij stelt het proleem dat binnenkomende informatie (talig of perceptueel) vaak ambigu is. Om die informatie te disambigueren moet je een aantal relevante feiten weten. Hoe weet je welke feiten relevant zijn? Dat weet je als je de context weet. Hoe weet je wat de context is? Daarvoor moet je weer een aantal relevante feiten weten enz. Men moet òf stellen dat een aantal feiten altijd relevant is en een vaste betekenis heeft, ongeacht de context - maar die mogelijkheid is uitgesloten toen een beroep op context juist nodig bleek - òf er is sprake van een regressie van contexten en manieren om ze te herkennen. Misschien stopt die regressie wel, misschien is er een uiteindelijke, breedste context. Maar die context zou gevormd worden door alles wat we weten, door onze hele leefwereld. Ons probleem was echter nu juist dat in principe alles wat we weten relevant kan zijn voor de beoordeling van een stukje binnenkomende informatie, maar dat we zochten naar een manier om de relevante informatie uit de oneindige hoeveelheid kennis te isoleren (zie Dreyfus 1979, 213-224) (26).

Het *frame*-probleem is in de AI niet onbesproken gebleven (27). Zo spreekt Minsky van *frames* in verband met de veranderende interpretatie van visuele informatie bij beweging in een stilstaande omgeving. Dit is een voorbeeld van het probleem van het bijwerken van oude informatie naar aanleiding van nieuwe informatie, namelijk de informatie dat men zelf van positie veranderd is (Minsky 1975a, 1975b). En Schank en Abelson spreken van *scripts:* een soort

scenario voor de normale gang van zaken in een bepaalde situatie. Deze *scripts* moeten een oplossing bieden voor het probleem om een hoeveelheid informatie te isoleren die in een bepaalde situatie relevant is (Schank en Abelson 1977). Maar deze voorstellen signaleren het probleem en geven het een naam; ze vormen geen oplossing voor het probleem. Zo heeft ieder script vele *pointers* naar andere scripts, zodat uiteindelijk alle scripts met elkaar verbonden zijn. Welke *pointers* moet je volgen en welke niet?

In de AI placht men dit soort problemen te zien als praktische problemen, een probleem van de organisatie en het bijhouden van de kennis, de *database* (28). Fodor gaat veel verder Hij verbindt de equipotentialiteit en niet localiseerbaarheid van centrale, niet-ingekapselde processen, met hun globaliteit en niet-modulariteit tot een sombere conclusie voor de AI Volgens hem zijn de grenzen van de modulariteit - bepaald door het criterium van de informationele inkapseling - ook de grenzen van wat we kunnen begrijpen van de geest met onze huidige theoretische midelen. De globale processen worden volgens hem noch in de cognitieve psychologie, waar men zoekt naar een formalisering van de centrale denkprocessen, noch in de wetenschapsfilosofie, waar men zoekt naar een formalisering van wetenschappelijke confirmatie, goed begrepen.

> "In this respect, cognitive science hasn't even *started*... If someone - a Dreyfus, for example - were to ask us why we should even suppose that the digital computer is a plausible mechanism for the simulation of global cognitive processes, the answering silence would be deafening" (Fodor 1983, 129).

In het kort niet alleen meent Fodor, in tegenstelling tot Pylyshyn, dat cognitieve ondoordringbaarheid geen criterium is om het niet-cognitieve af te grenzen van het cognitieve, en dat dus het cognitief ondoordringbare *ook* het domein vormt van AI en cognitieve psychologie, hij oppert tevens de mogelijkheid dat *alleen* het cognitief ondoordringbare het domein vormt van AI en cognitieve psychologie.

De mens-machine gelijkheid is in de cognitieve psychologie aangedragen als een empirische steun voor de voorgestelde fysicalistische oplossing van het lichaam-geest probleem. Het bestaan van mens-gelijke machines vormt immers een existentiebewijs voor de mogelijkheid dat op een volledig fysisch systeem twee soorten van beschrijvingen en verklaringen, in fysische termen en in intentionele termen, van toepassing zijn.

Maar we hebben gezien dat de mens-machine gelijkheid die de sterke AI propageert niet onproblematisch is. Aanvankelijk werd de gelijkheid gezien op het niveau van de kleinste informatiedragend eenheid: de alles-of-niets vurende neuron was functioneel gelijk aan de flip-flop, volgens de sterke AI. De vergelijking op dit fijnkorrelige niveau bleek niet houdbaar. Toen ging men zich concentreren op een gelijkheid op het niveau van het uiterlijk gedrag: mens en machine waren input-output equivalent volgens de sterke AI. De vergelijking op dit grofkorrelige niveau bleek evenwel niet vruchtbaar genoeg te zijn. Vervolgens deed Pylyshyn zijn methodologische voorstellen om te kunnen komen tot een sterke equivalentie van functionele architectuur en computationele processen bij mens en machine. Zijn voorstellen zouden leiden tot een vergelijking met middelfijne korrel. Tevens moest zijn criterium van cognitieve ondoordringbaarheid leiden tot een werkverdeling in de psychologie. Tenslotte liet Fodor zien dat die werkverdeling betwijfeld moest worden. Volgens hem vormen juist alleen de cognitief ondoordringbare processen een goed domein voor de AI. Daar zijn volgens hem de successen geboekt, in de psychologie en de AI van de input-systemen; en expertsystemen zijn alleen succesvol omdat ze informationeel ingekapseld zijn, slechts werkzaam op een nauwkeurig afgegrensd domein.

Voorspellingen over wat computers wel of niet zullen kunnen zijn erg gevaarlijk. De overenthousiaste uitspraken van AI-mensen uit de vijftiger en begin zestiger jaren zijn al voldoende belachelijk gemaakt (b.v. in Dreyfus 1972, 1979, McDermott 1976, 1981). Anderzijds lijkt soms de regel te gelden: "Zodra iemand zegt dat een computer iets nooit zal kunnen, komt er een programmeur die een machine precies dat laat doen" (b.v. Michie 1982). Over het algemeen is men in de

wereld van AI en cognitieve psychologie voorzichtiger geworden, het gevoel 'er bijna te zijn' heeft plaats gemaakt voor de overtuiging dat er nog vele moeilijkheden overwonnen moeten worden Fodor gaat verder wanneer hij beweert dat de grenzen van de modulariteit ook de grenzen zijn van wat we van de geest (*mind*) kunnen begrijpen " .given anything like the theoretical apparatus currently available" (Fodor 1983, 126) Volgens hem is er zelfs nog geen begin gemaakt met het begrijpen en formaliseren van de centrale denkprocessen Dreyfus gaat nog een stap verder volgens hem zijn de centrale, globale denkprocessen in principe niet formaliseerbaar, en dus niet programmeerbaar

Ik zelf vind Fodor s en Dreyfus' argumenten overtuigend en ben geneigd het met Dreyfus' conclusie eens te zijn, maar zal daar verder niet voor argumenteren (29) Niets in mijn betoog hangt namelijk af van de vraag of de moeilijkheden van het formaliseren en programmeren van centrale denkprocessen van praktische of van principiele aard zijn. Waar het mij om gaat is dat die moeilijkheden er zijn en dat een oplossing nog niet in zicht is Dat betekent dat er nu in elk geval *geen computers zijn die net als mensen zijn,* omdat ze een uiterst belangrijk deel van onze cognitieve processen, de globale, centrale denkprocessen, niet hebben We hadden trouwens ook al gezien dat er geen computers zijn die de Turingtest (voor enkel input-output-equivalentie) kunnen passeren En dat betekent dat de fysicalistische oplossing van het lichaam-geest probleem die gangbaar is in de cognitieve psychologie nu in elk geval niet kan steunen op het feitelijk bestaan van computers die net als mensen zijn. In tegendeel! Wanneer men gelooft dat de problemen bij het formaliseren en programmeren van praktische en niet van principiele aard zijn, dan is dat omdat men op andere gronden gelooft in die fysicalistische oplossing van het lichaam-geest probleem Men redeneert dan "De problemen moeten op de een of andere manier oplosbaar zijn, al weten we nu nog niet hoe, want de mens is toch ook een volledig fysisch systeem, en moet als zodanig nagebouwd kunnen worden"

Dit is een belangrijk punt We hebben gezien in 1 2 dat in de cognitieve psychologie het argument voor een fysicalistische oplossing van het lichaam-geest probleem steunde op een empirische en een apriori poot De empirische poot was dat computers net als mensen zijn

voor wat betreft de cognitieve processen Niet alleen in de cognitieve *psychologie* wordt er gewezen op die empirische poot Ook filosofen wijzen op het empirisch succes van de cognitiewetenschap ter ondersteuning van de filosofische theorie die ermee samengaat Die filosofische theorie expliciteert onder andere de fysicalistische oplossing voor het lichaam-geest probleem. Wanneer de hele cognitiewetenschap gezien wordt als een researchprogramma in de zin van Lakatos (1970), dan wordt de empirische progressie gezien als steun voor de harde (filosofische) kern van het programma Zo zegt Dennett over die filosofische harde kern, het functionalisme:

> "The burden for functionalism is inseparable from the burden of the variety of cognitive theories for which it provides the conceptual underpinnings" (Dennett 1978a, 256).

En Fodor beweert over het functionalisme (ofschoon hij later, in zijn boek over modulariteit, pessimistisch is over het programmeren van globale denkprocessen:

> "... there is this much to be said in its favor: It legitimizes the notion of mental representation, which has become increasingly important to theorizing in every branch of the cognitive sciences ... the science of mental representation is now flourishing" (Fodor 1981b, 132).

Mijn conclusie van dit hoofdstuk 2 luidt evenwel dat juist op het gebied van de 'hogere' cognitieve processen, de globale, centrale denkprocessen, er *geen* empirische progressie is in de AI. Het geloof dat de problemen op dit gebied slechts van tijdelijke en praktische aard zijn, berust op de filosofische theorie achter de cognitiewetenschap, en kan dus slechts op straffe van circulariteit gebruikt worden om die theorie te steunen

Nu de empirische ondersteuning van de fysicalistische oplossing van het lichaam-geest probleem bij nader onderzoek geen stand blijkt te houden, moet alle ondersteuning komen van een apriori argumentatie. De fysicalistische oplossing luidde. "De mens is een volledig fysisch systeem waarvan het gedrag en de werking zowel in louter fysische

termen alsook in intentionele termen beschreven en verklaard kan worden". De stroming in de *philosophy of mind* die ontstaan is in aansluiting op en ter fundering van de cognitieve psychologie en de AI werkt deze oplossing verder uit. Die filosofische theorie moet expliciteren wat de relatie is tussen beide soorten verklaringen en beschrijvingen, en welke eigenschappen van een fysisch systeem beide soorten verklaringen en beschrijvingen mogelijk maken. Zo'n theorie kan dan de sterke AI, de stelling dat computers onze mentale toestanden en processen echt hebben, legitimeren. Ze kan ook de hoop legitimeren dat de problemen bij het programmeren van globale denkprocessen zullen worden opgelost. Maar ze kan zelf geen steun ontlenen aan het bestaan van computers - zuiver fysische systemen - die al onze cognitieve processen hebben. Zulke computers bestaan er niet.

3 COGNITIEVE PSYCHOLOGIE EN FILOSOFIE VAN HET MENTALE

3 1 Inleiding

In hoofdstuk 1 hebben we gezien dat de cognitieve psychologie claimt een fysicalistische oplossing te kunnen bieden voor het lichaam-geest probleem Die oplossing leek te berusten op een empirisch argument en een apriori argument Zojuist, in hoofdstuk 2, hebben we die empirische poot nader bekeken Daar bleek dat de geclaimde oplossing geen empirische steun ondervond van het bestaan van computers In dit hoofdstuk wil ik de filosofische theorie bespreken die de fysicalistische oplossing voor het lichaam-geest probleem expliciteert De filosofische theorie die aansluit bij de cognitieve psychologie, en haar claims moet legitimeren, is het *functionalisme.* Om het functionalisme goed te kunnen kenschetsen, moet het gecontrasteerd worden met twee eerdere filosofische stromingen: het *behaviorisme* en de *identiteitstheorie.* In dit hoofdstuk zal ik de algemeen aanvaarde stellingen en argumenten van het functionalisme bespreken. Er wordt beargumenteerd waarom het behaviorisme en de identiteitstheorie onbevredigend zijn, en er wordt uiteengezet wat het functionalisme daar tegenoverstelt. Voorts worden de problemen van het functionalisme met betrekking tot Turingmachine-toestanden en die met betrekking tot qualia besproken. In de volgende hoofdstukken worden dan twee verschillende versies van het functionalisme uitgewerkt die als grondslag van en legitimering voor de cognitiewetenschap en de bijbehorende fysicalistische oplossing voor het lichaam-geest probleem kunnen dienen

3 2. Het behaviorisme.

In de eerste decennia van deze eeuw is het radicale behaviorisme opgekomen in de psychologie. Volgens deze stroming bestaat gedrag uit observeerbare responsen op observeerbare stimuli, en meer is er niet in verband met mensen en dieren; dat wil zeggen dat er geen mentale toestanden en processen bestaan, enkel fysische. Volgens een meer

gematigde versie, het methodologisch behaviorisme, was er misschien wel meer - iets mentaals - maar daar viel wetenschappelijk niets over te zeggen Er was onder de behavioristen een grote afkeer van alles wat 'mentalistisch' was Deze afkeer had te maken met het grote streven van de behavioristische psychologie naar 'wetenschappelijkheid' en het vermijden van alles wat met metafysica te maken leek te hebben, beide overeenkomend met het logisch positivisme van de Wiener Kreis (30). Men wilde niet werken met oncontroleerbare 'mentale entiteiten'. De verklaring van intelligent gedrag die de psychologie geacht wordt te geven mag niet *question begging* zijn Zo mag ze niet intelligentie verklaren in termen van intelligentie, door bijvoorbeeld slimme homunculi te postuleren aan de controlepanelen van het zenuwstelsel (zie Skinner 1964, Dennett 1978b).

Er bestond in het behaviorisme geen lichaam-geest probleem, geen vraag hoe iets mentaals ooit fysisch gedrag kan veroorzaken, want volgens het behaviorisme bestond het mentale niet, of was wetenschappelijk niet te benaderen Het behaviorisme in de psychologie streefde ernaar wetmatige relaties te ontdekken tussen stimuli en responsen, welke beide in fysicalistische termen uitgedrukt moesten worden. Men meende geen verklaringen en beschrijvingen in mentale termen nodig te hebben, hetzij omdat men meende dat het mentale überhaupt niet bestond, hetzij omdat men meende dat als het mentale bestond, het enkel een epifenomeen was dat er verder niets toe doet. Wanneer de relatie tussen stimulus en respons niet eenvoudig uit te drukken was, ging men er in de behavioristische psychologie soms toe over om een interveniërende variabele te postuleren Deze variabele diende echter uitsluitend om een mathematisch verband tussen stimulus en respons te kunnen uitdrukken, en mocht dan ook niet geïnterpreteerd worden als verwijzend naar een echte innerlijke toestand

Ook in de filosofie is een behaviorisme ontstaan, het *logisch behaviorisme.* Dit logisch behaviorisme is niet zozeer een methodologisch voorschrift om mentale termen in de psychologie te vermijden, maar een semantische theorie over de betekenis van mentale termen. Door een betekenisanalyse meende men het lichaam-geest probleem te kunnen opheffen In 1949 wilde Gilbert Ryle, in zijn *The concept of mind,* aantonen dat het hele probleem niet bestond Vóór

Ryle had er een *philosophy of mind* bestaan met allerlei '-ismes' zoals idealisme, materialisme, neutraal monisme, dualisme, epifenomenalisme, interactionisme, die alle een oplossing zochten voor het lichaam-geest probleem (in ruime zin). Ryle meende door 'zorgvuldige' conceptuele analyse te kunnen laten zien dat deze hele filosofie berustte op een grote vergissing, een *category mistake,* veroorzaakt door verwarrend en misleidend taalgebruik. Conceptuele analyse kon de verwarring verhelderen en het probleem als onzinnig terzijde schuiven.

Ryle probeert de notie van een mentale veroorzaking van gedrag te ridiculiseren. Hij bespreekt in *The concept of mind* (1949, 33 e.v.) de vraag "What makes a clown's clowning intelligent (witty, clever, ingenious, etc)?" Het antwoord dat hij afkeurt gaat als volgt: Wat het clownen intelligent maakt is het feit dat het het gevolg is van bepaalde mentale operaties (computaties, berekeningen) die alleen voor de clown toegankelijk zijn en die causaal verantwoordelijk zijn voor het produceren van het gedrag van de clown. Waren die operaties anders geweest (zo gaat dit antwoord verder), dan zou het clownen niet intelligent zijn geweest, of intelligent op een andere manier. Volgens Ryle kan zo'n antwoord niet juist zijn. Wat volgens Ryle het clownen echt knap maakt is iets heel anders: bijvoorbeeld, het feit dat de dingen die de clown doet niet de dingen zijn die het publiek verwachtte; het feit dat de man die hij met de slagroomtaart trof avondkleding droeg; het feit dat het gebeurt waar het publiek het zien kan enz. Dit soort feiten zijn helemaal niet privé voor de clown: het zijn juist de publieke aspecten van het gedrag die het clownen knap maken. En bovendien, wat het clownen knap maakt zijn niet de oorzaken van het gedrag, een knappe *ghost in the machine,* maar de aard van het gedrag zelf. Wat het clownen zo knap maakt is niet het feit dat het een gevolg is van een bepaald soort oorzaak.

Eenzelfde redenering past Ryle toe op de psychologie van perceptie: wat iets maakt tot de perceptie van een roodborstje is niet het voorkomen van een mentale gebeurtenis, maar het feit dat wat voor een roodborstje aangezien werd inderdaad een roodborstje was. Volgens Ryle maken mentalistische theorieën *category mistakes* omdat ze wat eigenlijk een *conceptuele* relatie is tussen verschillende aspecten van een enkele gebeurtenis behandelen als een *causale* relatie tussen twee verschillende gebeurtenissen. Zo bezien kan het mentale nooit een

oorzaak zijn van gedrag, het is veeleer een aspect van het gedrag De relatie tussen een mentale of psychologische toestand en gedrag is niet causaal maar conceptueel

"Ik nam aspirine omdat ik hoofdpijn had" lijkt te betekenen dat er iets in mij was, iets mentaals, dat mijn aspirine-neem-gedrag veroorzaakte, maar zo is het niet Er is geen causale relatie tussen hoofdpijn en aspirine-neem-gedrag, hoofdpijn staat in een conceptuele relatie tot bepaalde soorten gedrag, zoals aspirine nemen, kreunen, en zeggen· "Ik heb zo'n hoofdpijn" enz.

Het lijkt een probleem dat ik geen gedrag hoef te vertonen en toch hoofdpijn kan hebben. Dennett (1978b) geeft hiervan een gruwelijke illustratie. een tijdlang, in de veertiger jaren van deze eeuw, meende men dat de stof *curare* behalve een verlammende ook een anesthetische werking had Patiënten die onder curare werden geopereerd, waren volstrekt rustig onder het mes, maar klaagden achteraf bitter dat ze volledig bij kennis waren geweest en afgrijselijke pijnen hadden geleden. Ze hadden, tijdens de operatie, geen enkel gedrag vertoond Maar stel dat de patienten vóór de operatie met de curare ook nog een vergeetstof hadden gekregen die hun geheugen van de operatie achteraf uitwist Dan zou er helemaal geen pijngedrag zijn, ook niet achteraf (31). Een rechtgeaard behaviorist zou geen bezwaar mogen hebben tegen zo'n operatie: volgens hem zou hij geen pijn hebben (immers, fysiologische processen, die niet door de curare onderdrukt worden, zijn voor een behaviorist irrelevant voor de betekenis van mentale termen). Ik denk dat weinigen zo ver zouden gaan in hun filosofische overtuiging!

Maar de logisch behaviorist kan zeggen dat mentale termen, zoals hoofdpijn, niet logisch equivalent zijn met gedrag maar met gedrags-*disposities*. Ryle spreekt van (semi)hypothetische uitspraken (Ryle 1949), bijvoorbeeld: "Als er geen verlamming was geweest, dan was er pijngedrag opgetreden". Zulke (semi)hypothetische uitspraken kunnen wellicht nooit een sluitende definitie geven van mentale termen, omdat het vaak onmogelijk is om een opsomming te geven van alle hypothetische situaties waarin een gedragsdispositie tot uiting kan komen. Maar dat neemt voor de logisch behaviorist niet weg dat mentale termen logisch equivalent zijn met gedragsdisposities, en dat er geen zelfstandige mentale toestanden of gebeurtenissen bestaan als

oorzaken van gedrag.

3.2.1. Kritiek op het behaviorisme

Tegen de verschillende vormen van het behaviorisme zijn een aantal overtuigende argumenten ingebracht. Ik bespreek achtereenvolgens 1) de kritiek op het methodologisch behaviorisme in de psychologie, 2) de kritiek op Ryle's ridiculisering van mentale veroorzaking en 3) de kritiek op de definitie van mentale termen in termen van gedrag.

1) Het behaviorisme in de psychologie is voornamelijk een methodologisch voorschrift om mentale termen te vermijden en te zoeken naar S-R verbanden Naarmate de psychologie meer van zulke verbanden zou ontdekken zou het duidelijk worden dat het gedrag verklaard kan worden zonder mentale toestanden te postuleren. Het sterkste argument hiertegen is dat de psychologie *niet* meer en meer S-R verbanden ontdekte. Aanvankelijk leek het velen wel verstandig om bescheiden te beginnen en zich te concentreren op het leergedrag van ratten of het leren van nonsens-lettergrepen. Maar de S-R psychologie bleek ook na verloop van tijd niet of nauwelijks extrapoleerbaar naar meer complexe gedragingen (b v. Chomsky 1959). Psychologen ondervonden de methodologische voorschriften van het behaviorisme als verstikkend en onvruchtbaar.

2) Ryle's ridiculisering van de mentale veroorzaking van gedrag is vooral door Fodor bekritiseerd In de inleiding van zijn *The language of thought* (1975) geeft hij een zeer duidelijk en beroemd geworden weerlegging van Ryle's logisch behaviorisme, voor zover dit opgevat kan worden als een aanval op mentalistische psychologie (32) Fodor wil laten zien dat een mentalistische psychologie, met causale verklaringen van gedrag, wel mogelijk is.

Fodor weerlegt Ryle's aanval op mentalistische theorieën aan de hand van een ander voorbeeld dan het clownen. Hij bespreekt de vraag: "Wat maakt Wheaties het ontbijt van kampioenen?" Fodor leest deze vraag als "Wat in Wheaties maakt kampioenen van (sommige, zo vele) Wheaties-eters?" en niet als "Wat in Wheaties maakt dat (sommige, zo vele) kampioenen het eten?" Deze laatste versie vraagt om de redenen die de kampioenen geven voor het eten van Wheaties. Dat kan zijn

omdat het hun prestatie verhoogt, maar ook gewoon omdat ze het lekker vinden Een goed antwoord op de eerste versie van de vraag zou zijn "Wat Wheaties maakt tot het ontbijt van kampioenen is het aantal vitaminen en mineralen dat het bevat", of "De koolhydraten in Wheaties geven extra energie" of iets dergelijks. Waar het om gaat is dat dit een *causaal* verhaal is de antwoorden proberen de eigenschappen van Wheaties te specificeren die een causale rol spelen in de processen die van Wheaties-eters kampioenen maken De antwoorden zoeken een waarde voor P in het verklaringsschema 'P veroorzaakt ((x eet Wheaties) veroorzaakt (x wordt kampioen)) voor een significant aantal x' Dat is een causaal antwoord op de vraag die in de slogan gesteld wordt. Er is echter ook een heel ander antwoord mogelijk op de vraag· "Wat maakt Wheaties het ontbijt van kampioenen?" Dit tweede soort antwoord is helemaal niet causaal en stelt eenvoudigweg: "Wat Wheaties het ontbijt van kampioenen maakt is het feit dat een belangrijk aantal kampioenen het voor hun ontbijt eten" Dit is geen causaal maar een conceptueel antwoord; het noemt de conceptueel voldoende en noodzakelijke voorwaarde waaronder iets het ontbijt van kampioenen genoemd kan worden. De antwoorden die behoren bij het conceptuele verhaal horen niet bij het causale verhaal en *vice versa*. Het feit dat veel kampioenen het eten is er niet de *oorzaak* van dat Wheaties het ontbijt van kampioenen is, het behoort tot de analyse van het ontbijt van kampioenen zijn' Het gaat hier om een conceptueel verband.

Los van de vraag hoe de adverteerders hun slogan over Wheaties bedoeld hebben (33), toont het voorbeeld aan dat zowel een causaal als een conceptueel verhaal tegelijkertijd ware, verschillende antwoorden kunnen vormen op de vraag. 'Wat maakt (een) x (een) F? (wat maakt des clown's clownen knap?', wat maakt Wheaties het ontbijt van kampioenen?') Algemeen gezegd, stel dat C (onverwachtheid, gegeten worden door kampioenen) een conceptueel voldoende voorwaarde is voor het hebben van eigenschap F (knap, het ontbijt van kampioenen), en stel dat een individu a (des clown's clownen, Wheaties) voldoet aan C, zodat Fa een uitspraak is die waar is voor a. Dan is het daarmee normaliter helemaal niet uitgesloten om te vragen naar een causaal/mechanistische verklaring voor het feit dat Fa waar is Zo'n verklaring vormt een mogelijk antwoord op de vraag 'Wat maakt a F?'

Het feit dat a voldoet aan C vormt normalerwijze geen causaal/mechanistische verklaring van het feit dat a de eigenschap F vertoont, hoewel het feit dat a aan C voldoet een ander soort antwoord vormt op de vraag 'wat maakt Fa waar?'.

Fodor heeft hiermee willen laten zien dat een mentalistische psychologie wel mogelijk is, of liever, hij wil vooral laten zien dat het methodologisch vlekkeloos juist is om te zoeken naar theorieën over de gebeurtenissen die causaal mediëren in het produceren van intelligent gedrag. Dat het mogelijk is hoeft hij niet aan te tonen: cognitieve psychologen houden zich immers al bezig met dergelijke theorieën. Fodor wil laten zien dat ze geen *category mistake* maken, als ze zeggen dat mentale toestanden een causale rol spelen in het teweegbrengen van gedrag.

3) Op de mogelijkheid van vertaling van mentale termen in termen van gedrag is veel kritiek geleverd. De meeste kritiek richt zich op het feit dat mentale termen van toepassing kunnen zijn zonder dat er het bijpassende gedrag is, of dat men bepaald gedrag kan vertonen zonder dat de bijpassende mentale termen van toepassing zijn - de haast spreekwoordelijke clown uit vele liederen wiens onpeilbare droefheid schuilgaat achter zijn altijd vrolijke gedrag: "Niemand kende de pijn, van zijn stille verdriet" (Ben Cramer, "De clown") of "He's drowning his sorrow in whisky and gin" (Dave Davies, "Death of a clown"). Maar die kritiek is te ondervangen met een beroep op (semi)hypothetische uitspraken en *counterfactuals: "Als* de clown niet altijd geacteerd had, *dan..."*

Steekhoudender is de kritiek op het niet-episodische karakter van een aantal mentale termen. Volgens Ryle duiden uitspraken als 'hij besteedt aandacht aan zijn werk' of 'zij heeft plezier in haar werk' niet op een actuele mentale toestand of gebeurtenis; er wordt geen episode mee aangeduid. Het 'aandacht besteden aan' of 'plezier hebben in' staat enkel voor aspecten van het huidige en mogelijk toekomstige gedrag, het is niet iets nu naast mijn gedrag. Ryle spreekt ook liever van *met aandacht* of *met plezier* iets doen. Maar je kunt ook aandacht besteden aan andermans gedrag en plezier hebben in niets doen. Dan lijkt er toch sprake te zijn van een momentane mentale toestand. Ryle kan niet zeggen dat je kijkt met aandacht, want volgens hem is kijken zelf een vorm van aandacht besteden (zie Place 1967). En het plezier hebben in

het nietsdoen is wel degelijk een momentane toestand, dat merk je als je er plotseling geen plezier meer in hebt Je kunt de duur van je plezier in dat niets doen aangeven (Penelhum 1967, zie ook Marres 1985)

Het zwaarste argument tegen de mogelijkheid van vertaling van mentale termen in termen van gedrag is dat mentale termen nooit alleen voorkomen Om terug te komen op het voorbeeld van hoofdpijn hoofdpijn is niet het enige dat medieert tussen het niet functioneren van de airconditioning en mijn innemen van aspirine Er is ook sprake van mijn *kennis* van het bestaan van de eigenschappen van aspirine, mijn *mening* over geneesmiddelen, mijn *wens* om van mijn hoofdpijn af te raken In plaats van een mentale toestand die medieert tussen stimulus en respons, blijken er een heleboel mentale toestanden en processen te zijn, die met elkaar interacteren en waardoor ik uiteindelijk aspirine neem

Chisholm (1957) laat zien dat in de definitie van een mentale term altijd weer mentale termen moeten voorkomen Wanneer gelooft iemand in het bestaan van eenhoorns? Als hij de dispositie heeft om het geluid "Er bestaan eenhoorns" te maken Maar dat gaat alleen op als hij met het geluid "eenhoorns" eenhoorns *bedoelt*, naar eenhoorns *verwijst*. Wanneer verwijst iemand met het geluid "eenhoorn" naar eenhoorns? Als hij in aanwezigheid van eenhoorns op het geluid 'eenhoorn?" een bevestigende respons geeft Maar dat gaat alleen op als hij de eenhoorn *voor* een eenhoorn *houdt* en met zijn bevestigende respons een bevestiging *bedoelt*. Ook Davidson (1970) ontwikkelt een soortgelijk argument En Stegmuller (1974) laat de interdependentie tussen 'willen' en geloven' zien We blijven bij een poging tot definitie van mentale termen steeds gevangen in de kring van mentale termen (of van intensioneel taalgebruik a la Chisholm) (34)

Het behaviorisme vindt tegenwoordig nog maar weinig aanhang In de psychologie wordt niet langer getracht om mentale termen te vermijden, en in de filosofie is de notie van mentale veroorzaking van gedrag niet langer apriori verdacht

In de filosofie kwam, eind vijftiger jaren, een andere stroming in opkomst, sterk gericht op de empirische wetenschappen de identiteitstheorie Ook van deze theorie bestaan verschillende varianten, maar de voornaamste stelling die ze allemaal gemeen hebben is mentale toestanden en processen zijn identiek aan neurofysiologische toestanden en processen (Place 1956, Feigl 1958, Smart 1959, Armstrong 1968) Mijn hoofdpijn is niet alleen een gevolg of een bijverschijnsel van een bepaald hersenproces, mijn hoofdpijn *is* een hersenproces Natuurlijk wist men al heel lang dat mentale processen van alles te maken hebben met hersenprocessen, maar deze theorie sprak heel expliciet van identiteit De identiteitstheorie moest een empirische theorie worden, conceptueel voorgeschetst door de filosofie, maar met de details ingevuld door de wetenschappen Haar ontologie bevatte uitsluitend wetenschappelijk respectabele entiteiten - geen *elan vital*, geen psi-krachten, geen ectoplasma, geen onstoffelijke geest, maar alleen hersencellen en hun biochemische en fysiologische kenmerken

De identiteitstheorie moest een empirische theorie zijn, geen semantische theorie over de betekenis van mentale termen De verdedigers van de theorie stelden heel expliciet dat de identiteit die zij claimden contingent was Natuurlijk betekent hersenproces b' niet hetzelfde als 'neiging tot zelfmoord' of 'nabeeld' (Smart 1959) Het 'is' van de identiteit is geen 'is' van definitie (Place 1956) Men verhelderde de positie door analogieen aan te wijzen met bekende identiteiten zoals de identiteit van een bliksemschicht met elektrische ontladingen, de Morgenster met de Avondster, genen met DNA-moleculen

Enerzijds sprak men van identiteit uit overwegingen van zuinigheid als er zoveel prachtige correlaties zijn tussen mentale toestanden en processen en hersentoestanden en -processen - en de neurowetenschappen blijven steeds nieuwe correlaties aandragen - waarom spreken we dan van twee processen of toestanden in plaats van een? Problemen van wederzijdse causale beinvloeding zijn dan ook meteen opgelost die hersenprocessen en -toestanden die identiek zijn aan mentale processen en toestanden passen gewoon in de causale

keten van fysische gebeurtenissen En anderzijds was daar de computer, net in opkomst. De computer had zeker geen aparte ziel of geest - hij is immers door mensenhanden gemaakt, een machine. En toch vertoont de computer allerlei gedrag dat door mentale processen gemedieerd lijkt te zijn: schaakspelen, vertalen, patroonherkennen, probleemoplossen. De computer leek bij uitstek een voorbeeld van een volledig fysisch systeem dat toch mentale toestanden en processen had.

Zoals we al hebben gezien was men in de jaren vijftig en zestig zeer optimistisch over de mogelijkheden van de computer - in 1958 kondigde Simon aan:

> "It is not my aim to surprise or shock you .. But the simplest way I can summarize is to say that there are now in the world machines that think, that learn and that create. Moreover, their ability to do these things is going to increase rapidly until - in a visible future - the range of problems they can handle will be coextensive with the range to which the human mind has been applied" (Simon en Newell 1958, 6).

Bovendien was men erg onder de indruk van de overeenkomst tussen computers en hersenen (zie 2.3.1). Beide bestaan in laatste instantie uit elektrische elementen die slechts twee mogelijkheden hebben: vuren of niet-vuren van neuronen en aan-uit van de binaire computer flip-flops. Beide kunnen zeer grote aantallen van uitsluitend deze elementen combineren en zodoende intelligent gedrag vertonen. De computer verleende aan de identiteitstheorie veel steun, dacht men.

Voor psychologen overtrof de introductie van de computer hun stoutste dromen. Ze konden nu vrijelijk mentalistische theorieën ontwerpen zonder alle verstikkende restricties van het vroegere behaviorisme, terwijl ze toch veilig verankerd bleven in een materialistische filosofie. Mentale termen stonden niet, zoals Skinner vreesde, voor twijfelachtige interveniërende variabelen, *mental way stations* (Skinner 1964), maar voor bestaande hersen- of computerprocessen. De computer was, zo meende men, in alle relevante (fysiologische, niet chemische) aspecten gelijk aan de hersenen. En omdat volgens de identiteitstheorie de hersenen alle gedrag

veroorzaken, kunnen door computersimulatie alle vragen van de psychologie, de studie van het menselijk gedrag, worden beantwoord. Dat was de legitimatie voor computergebruik in de cognitieve psychologie. De psychologie kon aan het werk.

3.3.1. Kritiek op de identiteitstheorie.

De kritiek op de identiteitstheorie valt in drie soorten argumenten uiteen technische argumenten over de aard van computers, filosofische argumenten over de aard van identiteit, en empirische en wetenschapsfilosofische argumenten over de reduceerbaarheid van psychologie tot neurofysiologie en fysica.

De argumenten over de aard van computers zijn al vermeld in 2.3.1: de gelijkstelling van hersenen en computer op het fijnkorrelige niveau van de kleinste informatiedragende eenheid, het alles-of-niets vurende neuron en de digitale flip-flop, bleek niet houdbaar. Bovendien bleek het leervermogen van computers nogal tegen te vallen - ze konden lang niet alles wat mensen kunnen (zie ook 2.4) Deze steun voor de identiteitstheorie bleek geen stand te houden.

De argumenten over de aard van identiteit en de wet van Leibniz staan vermeld in 1.3 Ik ken mijn mentale toestand als bijvoorbeeld enthousiast, en een hersenonderzoeker kent diezelfde toestand (mentale toestanden zijn identiek met hersentoestanden volgens de identiteitstheorie) als bijvoorbeeld hoogfrequent vurend met een lage amplitude. Zijn dat niet twee heel verschillende eigenschappen, en kan een toestand wel beide eigenschappen hebben? Om de identiteit überhaupt te kunnen stellen heeft men toch een dualism van talen of wijzen van kennen of van eigenschappen nodig Om dit dualisme te vermijden gingen sommigen over tot de extremere positie van een eliminatief materialisme

De empirische en wetenschapsfilosofische argumenten aangaande de reduceerbaarheid van psychologie tot neurofysiologie zijn in 1.2 al even aangestipt maar zullen hier uitgebreid besproken worden. De argumenten draaien om het punt dat de identiteit van mentale toestanden met hersentoestanden op twee manieren uitgelegd kan worden: als *type-identiteit* en als *token-identiteit*. Men spreekt van

een type-identiteit, en van *type-fysicalisme*, wanneer de identiteitsclaim zo wordt uitgelegd dat iedere mentale toestand van een bepaald mentaal soort identiek is met een fysische toestand van een bepaald fysisch soort; bijvoorbeeld iedere gedachte over aspirine is identiek met hersenproces *a* met fysische kenmerken F, G, H enz Men spreekt van een token-identiteit, en van *token-fysicalisme*, wanneer de identiteitsclaim zo wordt uitgelegd dat iedere feitelijk voorkomende mentale toestand identiek is met een of andere fysische toestand, terwijl mentale toestanden van hetzelfde soort identiek kunnen zijn met fysische toestanden die verschillend zijn in hun fysische eigenschappen.

Oorspronkelijk werd de identiteitstheorie uitgelegd als een type-fysicalisme, maar dit type-fysicalisme werd algemeen veel te sterk bevonden (b.v. Davidson 1970, Fodor 1974 (herdrukt in 1981a), Dennett 1978b) De oorspronkelijke, type-identiteitstheorie had problemen met de generaliseerbaarheid Gisteren om drie uur dacht ik aan aspirine Volgens de identiteitstheorie is die gedachte identiek aan mijn hersenproces-om-drie-uur-gisteren *a* met fysische kenmerken F, G, H enz. Het is nauwelijks plausibel om te veronderstellen dat *iedere gedachte* over aspirine een hersenproces *a* is met precies die fysische kenmerken F, G, H enz Zouden niet ook wezens met een andere fysiologie dan de mijne over aspirine kunnen denken, bijvoorbeeld marsmannetjes, lichaamloze geesten, computers? Zo ook kan logische staat A van een Turingmachine in de ene fysische machine anders gerealiseerd zijn dan in de andere. De filosoof Hilary Putnam merkte in een nu als klassiek beschouwd artikel uit 1960, 'Minds and Machines', op, dat de vraag of men kan spreken van een strikte identiteit tussen de logische staat van een Turingmachine als wiskundige beschrijving en de concrete realisering ervan als fysische toestand in een werkelijke machine, onbeantwoordbaar was

Het is zelfs niet plausibel dat *iedere gedachte van mij* over aspirine identiek is aan dat hersenproces-om-drie-uur-gisteren *a* met fysische kenmerken F, G, H enz. Voortdurend sterven er neuronen af die niet meer bijgemaakt kunnen worden: ik kan morgen niet meer een hersenproces hebben identiek aan proces *a* - en dan hebben we het nog niet eens over grote lesies in de hersenen of Lashley's equipotentialiteitsthese, die zegt dat voor vele functies enkel een

bepaalde hoeveelheid hersencellen nodig is, ongeacht waar die zich bevinden. (Bij experimenten was hem gebleken dat vele taken niet te lijden hadden onder grote hersenlesies, ongeacht waar die werden toegebracht, zolang er een bepaald percentage hersencellen maar intact bleef) En bovendien, welke van de fysische kenmerken F, G, H enz maakt proces *a* tot een gedachte, en nog wel een gedachte over aspirine?

De identiteitstheorie werd in de sterke zin van een type-identiteit uitgelegd omdat men aanvankelijk meende dat fysicalisme, de leer dat alle entiteiten en gebeurtenissen fysische entiteiten en gebeurtenissen zijn, onverbrekelijk verbonden was met het reductionisme, de leer dat alle empirische wetenschappen gereduceerd kunnen worden tot fysica. Jerry Fodor laat in zijn artikel 'Special sciences' uit 1974 (herdrukt in Fodor 1981a) zien dat het standaard-reductionisme veel te sterk is, maar dat een andere relatie tussen de speciale wetenschappen en de fysica mogelijk is die compatibel is met het token-fysicalisme. Ik geef zijn argumenten voor een fysicalisme zonder (standaard-)reductionisme kort weer.

De standaard reductie gaat als volgt:

Laat formule 1) een wet zijn van de speciale wetenschap S.

1) $S1x \longrightarrow S2y$.

In vertaling: alle gebeurtenissen die eruit bestaan dat x is S1 leiden tot gebeurtenissen die eruit bestaan dat y is S2. S1 en S2 zijn typische predicaten van wetenschap S. Voor de reductie van deze wet zijn de volgende formules nodig:

2a) $S1x \longleftrightarrow P1x$

2b) $S2y \longleftrightarrow P2y$

3) $P1x \longrightarrow P2y$

P1 en P2 zijn predicaten van de fysica, en formule 3 is een wet uit de fysica. De formules 2a en 2b worden de brugformules genoemd De reductie van wetenschap S vereist dat ieder predicaat dat verschijnt in een wet van S als antecedent of consequent, moet verschijnen in een brugformule.

Om openingen voor dualisme te voorkomen, nemen veel filosofen aan dat de brugformules zoals formules 2a en 2b gezien moeten worden als identiteit van gebeurtenissen: iedere gebeurtenis die eruit bestaat dat x P1 is, is identiek aan een gebeurtenis die eruit bestaat dat x S1 is.

Deze lezing garandeert het fysicalisme, immers, zo is iedere gebeurtenis die onder een wet van een wetenschap valt een fysische gebeurtenis, die onder een wet van de fysica valt.

Het fysicalisme stelt dat de brugformules een contingente identiteit van gebeurtenissen uitdrukken. De zwakke vorm van het fysicalisme, token-fysicalisme, stelt dat iedere *gebeurtenis* die genoemd wordt in een wetenschappelijke wet een fysische gebeurtenis is. Een sterkere vorm, het type-fysicalisme, stelt dat iedere *eigenschap* die genoemd wordt in de wetenschappen een fysische eigenschap is. De gebeurtenis die bestaat uit het zetten van mijn handtekening was in een bepaald geval identiek aan de gebeurtenis die bestaat uit het kopen van een huis. Maar (godzijdank) is niet de eigenschap van het zetten van mijn handtekening identiek aan de eigenschap van het huis-kopen.

Nu is het zo dat iedere wetenschap de gebeurtenissen op het eigen terrein op een bepaalde manier indeelt. Een wetenschap heeft idealiter een vocabulaire van predicaten, zodat de gebeurtenissen onder de wetten van die wetenschap vallen voor zover ze voldoen aan de predicaten. Niet iedere ware beschrijving van een gebeurtenis is een beschrijving in het vocabulaire van die wetenschap; niet iedere ware beschrijving van een gebeurtenis beschrijft die gebeurtenis als instantie van een *soort* in die wetenschap. Gebeurtenissen die voor de éne wetenschap tot dezelfde soort behoren, kunnen voor de andere wetenschap tot verschillende soorten behoren.

Fodor (1981a, 133) noemt drie redenen waarom het onwaarschijnlijk is dat iedere soort van alle wetenschappen correspondeert met een fysische soort: a) er kunnen vaak interessante generalisaties gemaakt worden over gebeurtenissen waarvan de fysische beschrijvingen niets gemeen hebben, b) de kwestie *of* de fysische beschrijvingen van de gebeurtenissen die onder zo'n generalisatie vallen iets met elkaar gemeen hebben is vaak totaal irrelevant voor de waarheid van de generalisatie, of de interessantheid ervan, of de graad van confirmatie, of voor enige epistemologisch belangrijke eigenschap van de generalisatie; en c) de speciale wetenschappen houden zich juist bezig met dit soort generalisaties.

Neem bijvoorbeeld een wet uit de economie over het kopen van iets. Iedere koop heeft wel een ware beschrijving in het vocabulaire van de

fysica, en iedere koop valt dus onder de wetten van de fysica Maar niet allemaal onder dezelfde wet. Sommige kopen worden gesloten met handje-klap. Andere met het zetten van een handtekening. Weer andere misschien met het ritueel slachten van een kip. De fysische beschrijving die al die aankopen omvat is heel erg disjunctief. Zo'n disjunctie van fysische predikaten vormt dan de rechterkant van een brugformule. Maar de disjunctie vormt voor zover we weten geen natuurlijke soort. Wat is de kans dat die disjunctie de antecedent of de consequent vormt van een wet van de fysica?

Verschillende aankopen hebben iets interessants gemeen; de economische wet waarvan we het bestaan even hebben aangenomen zegt daar iets over. Maar wat interessant is aan aankopen van huizen is niet hun overeenkomsten onder een *fysische* beschrijving Bovendien, de economische wet gaat ook op voor alle *mogelijke* aankopen, waarvan we de fysische beschrijving nog helemaal niet kennen

De rechterkant van de meeste brugformules zal dus bestaan uit een disjunctie, en dan vaak nog een open disjunctie, dat wil zeggen een disjunctie met een onbepaald (of zelfs oneindig) aantal disjuncten. Om dit soort redenen is het onwaarschijnlijk dat de psychologie reduceerbaar zal zijn tot de neurofysiologie. Wat de irreduceerbaarheid van de psychologie garandeert is dat er geen één op één correspondenties zijn tussen psychologische en fysiologische soorten

Fodor legt er de nadruk op dat brugformules aan de rechterkant disjunctief zijn Daarmee laat hij zien dat type-fysicalisme en reductionisme in de gangbare zin niet houdbaar zijn. Het is niet zo dat iedere psychologische soort (type) gelijk is aan, of coëxtensief met, een neurofysiologische soort. Maar het is volgens hem nog wel steeds houdbaar dat iedere feitelijk voorkomende psychologische gebeurtenis een neurofysiologische gebeurtenis is. En daar zijn volgens hem ook empirische aanwijzingen voor te vinden. Stel we vinden dat voor ieder n-tal van soort-identieke psychologische gebeurtenissen, er een spatio-temporeel gecorreleerd n-tal is van *soort-verschillende* neurofysiologische gebeurtenissen Dat wil zeggen, iedere psychologische gebeurtenis gaat gepaard met een of andere neurofysiologische gebeurtenis, maar psychologische gebeurtenissen van dezelfde soort gaan soms gepaard met neurofysiologische

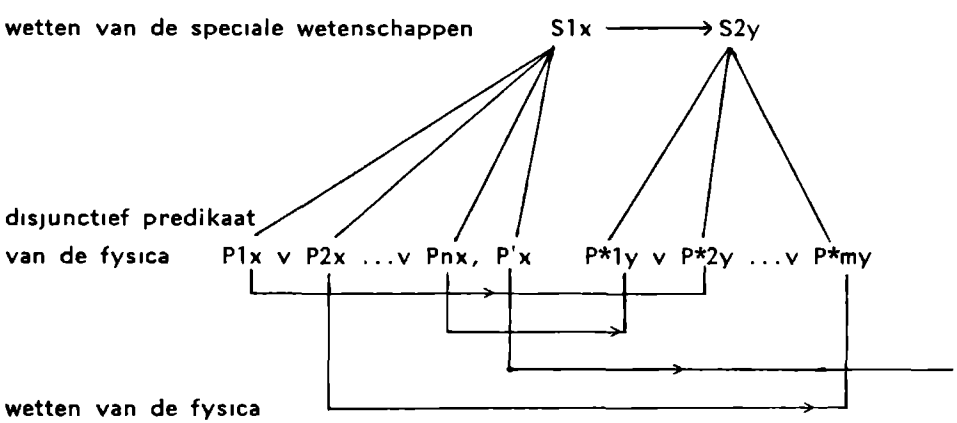
gebeurtenissen van verschillende soort. Token-fysicalisme vindt steun als we kunnen aantonen dat de verschillende neurofysiologische gebeurtenissen die gepaard gaan met een bepaald soort psychologische gebeurtenissen dezelfde psychologische eigenschappen hebben. De neurofysiologische gebeurtenissen verschillen dan voor wat betreft hun neurofysiologische eigenschappen, maar niet voor wat betreft hun gevolgen voor het gedrag. Het is immers niet zo dat type-verschillende gebeurtenissen geen enkele eigenschap gemeen hebben.

Volgens de standaard-versie van het reductionisme, het type-fysicalisme, zijn psychologische soorten type-identiek met neurofysiologische soorten, en bestaan er dus simpele brugformules. Maar volgens het token-fysicalisme kan iedere instantie van één en dezelfde psychische soort token identiek zijn met een instantie van telkens een andere neurofysiologische soort. De brugformules zijn dan disjunctief. In plaats van de standaard versie van het reductionisme stelt Fodor een andere versie voor van de relatie tussen de speciale wetenschappen en de fysica, die compatibel is met het token fysicalisme. In plaats van de volgende serie formules:

1) $S1x \longrightarrow S2y$
2a) $S1x \longleftrightarrow P1x$
2b) $S2y \longleftrightarrow P2y$
3) $P1x \longrightarrow P2y$

krijgen we nu het plaatje van figuur 3.

Figuur 3 (overgenomen uit Fodor 1981a, 139).



Er zijn, naast de compatibiliteit met het token-fysicalisme, nog twee voordelen verbonden aan deze constructie van de relatie tussen de wetenschappen: we kunnen nu toelaten dat wetten van de speciale wetenschappen uitzonderingen hebben, en we kunnen zien waarom er speciale wetenschappen zijn. Bij de oude constructie kon men niet verantwoorden dat de speciale wetenschappen uitzonderingen in hun wetten hebben maar de fysica niet Als P1x zonder uitzondering leidt tot P2y, dan moet ook S1x zonder uitzondering leiden tot S2y, tenzij de brugformules uitzonderingen kennen. Maar als je dàt toelaat ben je geen reductionist meer, want dan zouden er gebeurtenissen bestaan die geen fysische gebeurtenissen zijn. Met deze constructie is dat evenwel geen probleem· iedere instantiatie van S1 die contingent identiek is aan een instantiatie van P' vormt een uitzondering op S1x ⟶ S2y. Er is immers geen wet die P' verbindt met een P* predikaat. Er bestaat nu geen wet in de fysica die precies correspondeert met de wet S1x ⟶ S2y, alleen een aantal wetten, P1x ⟶ P*2y, P2x ⟶ P*my, enz., die ermee te maken hebben. Die wetten kun je wel verzamelen in de formule:

P1x v P2x v ... v Pnx ⟶ P1*y v P*2y v ...v P*ny,

maar dat is zelf geen wet Vergelijk: het is een wet, dat daling van temperatuur tot onder 0 graden Celsius leidt tot bevriezing van water, en het is een wet dat wrijving leidt tot warmte, maar het is niet een wet dat (òf daling van temperatuur tot onder 0 graden Celsius of wrijving) leidt tot (òf bevriezing van water òf warmte)

De identiteitstheorie als een type-fysicalisme dat claimt dat mentale soorten overeenkomen met neurofysiologische soorten, wordt niet veel meer aangehangen De soorten van de psychologie vormen vaak een dwarsdoorsnede van de soorten van de neurofysiologie, zo erkent men. De zwakkere vorm van het fysicalisme, het token-fysicalisme, dat stelt dat ieder feitelijk voorkomende mentale gebeurtenis een fysische gebeurtenis is, wordt wel aangehangen in de filosofie van het mentale; maar aanhangers van dit token-fysicalisme noemen hun theorie doorgaans geen identiteitstheorie, ook al beweren ze dat iedere mentale gebeurtenis identiek is met een fysische gebeurtenis.

De nieuwste filosofische theorie die een fysicalistische oplossing voor het lichaam-geest probleem tracht te expliciteren is het *functionalisme.* Het functionalisme sluit aan bij de kritieken op het behaviorisme en de identiteitstheorie. Anders dan het behaviorisme meent het functionalisme dat het zinvol is om in de psychologie mentale termen te gebruiken, dat het spreken over mentale oorzaken geen *category mistake* inhoudt, en dat mentale termen altijd in relatie tot elkaar gedefinieerd moeten worden. En anders dan de identiteitstheorie meent het functionalisme dat het type-fysicalisme veel te sterk is; alleen een token-fysicalisme acht men houdbaar

Nu is het token-fysicalisme een zeer bescheiden vorm van fysicalisme, en daardoor goed te verdedigen. De zwaarwegende argumenten, zowel empirische als wetenschapsfilosofische, die in te brengen zijn tegen het type-fysicalisme, gaan voor het token-fysicalisme niet op. Maar het token-fysicalisme zegt dan ook niet zoveel, met name beantwoordt het niet de vraag die het behaviorisme en het type-fysicalisme wel beantwoordden: "Wat hebben twee mensen (systemen) gemeen die dezelfde mentale toestand hebben?" Volgens het behaviorisme luidde het antwoord: "Ze hebben dezelfde gedragsdispositie", en volgens het type-fysicalisme luidde het antwoord: "Ze hebben eenzelfde neurofysiologische hersentoestand". Het token-fysicalisme heeft geen antwoord, volgens deze leer hoeven twee mensen (systemen) met dezelfde mentale toestand *niets* fysisch gemeen te hebben, noch in het gedrag, noch in de hersenen.

Het functionalisme geeft wel een antwoord op de vraag "Wat hebben twee mensen gemeen die dezelfde mentale toestand hebben, bijvoorbeeld aan aspirine denken?" Dat antwoord luidt. "Die twee mensen hebben beide een toestand die identiek is voor wat betreft de *functionele* eigenschappen ervan. Een mentale toestand of proces wordt gedefinieerd als die toestand of dat proces dat een bepaalde functionele rol speelt in relatie tot bepaalde stimuli, responsen en andere mentale toestanden en processen".

Het functionalisme stelt dat alles wat dezelfde rol speelt tussen kapotte airconditioning, hoofdpijn, meningen over geneesmiddelen, aspirine-inneem gedrag enz. een gedachte over aspirine *is.* Het

functionalisme is een theorie die het eerst voorgesteld is door Putnam (1960, 1964, 1975). Deze doctrine zegt op zichzelf niets over de feitelijke realisering van functioneel gedefinieerde mentale processen of toestanden Zo zegt het begrip 'muizeval' ook niets over de feitelijke realisering van zo'n ding, het is een functioneel gedefinieerd begrip. In principe laat het functionalisme de mogelijkheid open dat mentale toestanden en processen kunnen voorkomen bij engelen of geesten, en laat het zich combineren met vormen van dualisme. Maar gecombineerd met het token-fysicalisme, zoals het meestal wordt opgevat, stelt het functionalisme dat iedere mentale toestand of proces gerealiseerd wordt door een of andere fysische toestand of proces. Welke fysische eigenschappen zo'n toestand of proces heeft is dan niet van belang. Van belang zijn de functionele eigenschappen van die fysische toestand of dat proces Volgens de functionalist hebben de hersenen eigenschappen die in zekere zin niet fysisch zijn. Dat wil zeggen, eigenschappen die definieerbaar zijn in termen die geen melding maken van de fysica of biochemie of fysiologie van de hersenen.

Als het vreemd lijkt te spreken van een fysisch systeem met niet-fysiche eigenschappen, neem dan bijvoorbeeld een computer. Een computer heeft vele fysische eigenschappen. Hij heeft een bepaald gewicht, een bepaald aantal chips of transistoren enz. Hij heeft economische eigenschappen, zoals een bepaalde marktwaarde En hij heeft functionele eigenschappen, zoals het hebben van een bepaald programma Die laatste eigenschap is niet-fysisch in die zin dat hij gerealiseeerd kan worden in systemen van verschillende ontologische compositie. Een mens kan een bepaald programma realiseren, en een machine kan een bepaald programma realiseren, en de functionele organisatie van deze twee, de mens en de machine, kan exact hetzelfde zijn terwijl hun materiéle samenstelling totaal verschillend is. Volgens de functionalist hebben psychologische toestanden, net als programma's, functionele eigenschappen; dezelfde psychologische toestand, bijvoorbeeld boos zijn, kan een geïnstantieerd zijn in duizenden verschillende species die een heel verschillende fysiologie of biochemie kunnen hebben. Sommige van die species zouden buitenaards kunnen zijn; en misschien kunnen robots ooit boosheid vertonen.

Volgens het functionalisme zijn mentale toestanden neurofysiologische toestanden, die geïndividueerd worden door de rol die ze spelen ten

opzichte van input, output en andere mentale toestanden. Mentale toestanden zijn dus functionele toestanden.

Nu is de notie van een functionele toestand niet erg precies. Iedere toestand van ieder systeem kan gezien worden als een functionele toestand (zie Kalke 1969, Rorty 1972). Bovendien kan het postuleren van functionele toestanden om gedrag (of input-output relaties) te verklaren erg goedkoop zijn. Een functionele toestand heeft dan veel weg van een interveniërende variabele, behalve dat bij een functionele toestand verwezen wordt naar andere interne toestanden en bij een interveniërende variabele alleen naar input en output (Nelson 1976). Een functionalistische verklaring kan als volgt gaan: Wat veroorzaakt het spreken? Een spraakgenerator. En wat is een spraakgenerator? Alles wat de functionele rol speelt van het veroorzaken van het spreken.

De behaviorist Skinner was juist zo tegen dit soort *mental way stations,* omdat hij vreesde dat het zoeken naar verklaringen daar zou stoppen (Skinner 1964). Maar de functionalist wil functioneel gedefinieerde theoretische constructen alleen toelaten als er een mechanisme bestaat dat de functie kan uitvoeren, of als men mag aannemen dat zo'n mechanisme gebouwd kan worden. Het functionalisme stelt dus dat mentale toestanden gedefinieerd kunnen worden als functionele toestanden, en gerealiseerd worden in fysische systemen. Dat de fysische eigenschappen van zulke systemen (b.v. mens of machine) kunnen verschillen is niet relevant: de mentale toestanden vervullen dezelfde functionele rol en zijn dus hetzelfde.

In de volgende paragraaf zullen we de mogelijkheden bespreken om de notie van functionele toestand wat verder te systematiseren, en in 3.4.2 zal besproken worden of het functionalisme met alle soorten mentale toestanden even goed uit de voeten kan.

3.4.1. Functionele, logische en computationele toestand.

De functionalist wil functioneel gedefinieerde theoretische constructen alleen toelaten wanneer er een mechanisme bestaat dat de functie kan uitvoeren. Wanneer functionele toestanden gelijk gesteld worden aan de logische toestanden van een Turingmachine dan is aan deze eis

voldaan: iedere gespecificeerde Turingmachine kan fysisch gerealiseerd worden. Nelson (1976) stelt dan ook voor om 'functionele toestand' overal te vervangen door 'logische toestand'. Putnam stelt in zijn baanbrekend artikel 'Minds and machines' (1960) al dat mentale toestanden corresponderen met (in latere (1964, 1965, 1967) artikelen gelijk zijn aan) de logische toestanden van een Turingmachine. Een Turingmachine kan gekenschetst worden, zoals we zagen in 2.2 1, als een mechanisme met een eindig aantal logische toestanden. De input en de output van de machine zijn geschreven op een band die verdeeld is in vakjes waarop telkens één symbool van een eindig alfabet staat De machine 'leest' de band per vakje, kan een symbool uitwissen en een nieuw schrijven. De machine kan alleen de elementaire operaties verrichten van 'lezen', uitwissen, schrijven, band verschuiven en van logische toestand veranderen.

Nu is het mogelijk allerlei totaal verschillende systemen te beschrijven als dezelfde Turingmachine. Zo kan *alles* beschreven worden als een nul-Turingmachine: een machine met maar een logische toestand die geen output geeft, wat de input ook mag zijn. Ongeacht welk symbool op de band onder de leeskop schuift (de input), de machine blijft in dezelfde toestand, en verandert het symbool niet. Als je van een mens of dier of steen *niets* rekent als toestandsverandering en *niets* als output, dan heb je zo'n nul-Turingmachine. Het ligt er maar aan wat voor niveau van abstractie je gebruikt in je beschrijving. Je kunt de mens ook beschrijven als een Turingmachine met twee toestanden: dood of levend. Slechts bij zeer bepaalde inkomende stimuli vindt er een toestandsverandering plaats van levend naar dood. De omgekeerde toestandsverandering komt in de machinetabel niet voor. Je abstraheert dan van alle andere toestandsveranderingen die we gewoon zijn bij mensen te onderscheiden. Zo is het ook mogelijk om mijn kat als dezelfde Turingmachine te beschrijven als een muizeval (Kalke 1969). Maar natuurlijk is zo'n beschrijving van een twee-toestandige muizenvanger niet interessant vanuit het oogpunt van kattenpsychologie (Nelson 1976), net zo min als de twee-toestandige mens-machine (levend of dood) interessant is vanuit het oogpunt van mensenpsychologie. Het Turingmachine-functionalisme stelt dan ook dat er een unieke *beste* beschrijving (voor psychologische doeleinden) is van mensen en katten zodat hun mentale of psychologische toestanden

de logische toestanden van een Turingmachine zijn

Het Turingmachine-functionalisme is dus bestand tegen beschuldigingen van imprecisie en trivialiteit De Turingmachine is een duidelijk model voor een organisme, en psychologische toestanden zijn logische toestanden van die Turingmachine volgens de theorie. Maar er zijn een aantal bezwaren in te brengen tegen de opvatting dat psychologische toestanden corresponderen met of identiek zijn aan logische toestanden van een Turingmachine. Putnam heeft heel expliciet teruggenomen dat psychologische toestanden logische toestanden zijn van een Turingmachine (Putnam 1973) en Block en Fodor argumenteren in een artikel 'What psychological states are not' (1972, herdrukt in Fodor 1981a) eveneens dat psychologische toestanden geen logische toestanden zijn. Hun argumenten zijn overtuigend (zie ook Block 1979, Dennett 1978b). Ik geef ze kort weer

1) Het logische toestand functionalisme kan geen onderscheid maken tussen dispositionele toestanden en vóórkomende toestanden. Men zou kunnen zeggen dat een organisme in een voorkomende toestand is (bijvoorbeeld *nu* de gedachte heeft (voelt?) dat aspirine helpt tegen hoofdpijn), wanneer het *in* een bepaalde logische toestand a is, en dat het in een dispositionele toestand is (bijvoorbeeld de (al dan niet bewuste) gedragssturende mening heeft dat aspirine helpt tegen hoofdpijn), als diezelfde logische toestand a voorkomt in de machinetabel. Dat zou evenwel betekenen dat de mogelijkheid hebben om *p* uberhaupt te kunnen denken gelijk is aan het vóórkomen van een bepaalde logische toestand (a) in de machinetabel. Immers, al die toestand niet voorkomt in de tabel kunnen we volgens de theorie nooit *p* denken. Maar we hadden net gezegd dat het voorkomen van toestand a in de machinetabel gelijk was aan het dispositioneel menen dat *p*. Volgens deze constructie kunnen we alleen de gedachten hebben (in de zin van *entertainen)* die we dispositioneel menen, die ons gedrag sturen. Dat is absurd· ik ben van mening dat er geen eenhoorns bestaan, en mijn gedrag wordt vast niet onbewust gestuurd door de dispositionele mening dat ze wel bestaan, maar ik heb wel degelijk de mogelijkheid de gedachte te hebben dat ze bestaan, bijvoorbeeld bij het bedenken van dit voorbeeld (Fodor en Block)

2) Het Turingmachine-functionalisme heeft er moeite mee dat er meerdere psychologische toestanden tegelijk kunnen zijn. Neem een

bepaalde pijn als psychologische toestand Mijn logische Turing toestand moet niet alleen aangeven dat ik die pijn heb, maar ook of ik op het punt sta 'te' te schrijven, of ik een overvliegende straaljager hoor enz. De logische toestand waar ik nu in ben moet al dit soort dingen specificeren, die zeker niet allemaal behoren tot één psychologische toestand (Putnam). Mijn gedrag is het resultaat van een interactie tussen input en allerlei psychologische toestanden. De output van een Turingmachine is het resultaat van de interactie tussen input en allerlei logische toestanden. Maar die output kan alleen het resultaat zijn van een achtereenvolgende serie van logische toestanden, terwijl mijn gedrag ook het resultaat kan zijn van tegelijk voorkomende psychologische toestanden (Fodor en Block).

3) Logische toestanden hoeven helemaal geen kwalitatieve inhoud te hebben; omdat een logische toestand uitsluitend functioneel gedefinieerd is, dat wil zeggen in termen van input, output en andere logische toestanden, zou hij de functionele rol van bijvoorbeeld pijn kunnen spelen zonder enige pijnkwaliteit (Fodor en Block, zie ook 3.4.2).

4) Het Turingmachine-functionalisme construeert 'zelfde psychologische toestand' te nauw Twee organismen zijn in dezelfde psychologische toestand als ze in dezelfde logische toestand zijn, volgens de theorie. Maar ze zijn alleen in dezelfde logische toestand als hun volgende toestand ook gelijk is. En de volgende. En de daarop volgende. Dat zou betekenen dat twee organismen alleen in dezelfde psychologische toestand zouden verkeren als ze al hun psychologische toestanden exact gemeen hebben, èn al hun gedrag. Dit geeft een veel te fijnkorrelige indeling van psychologische toestanden (Fodor en Block).

5) Het Turingmachine-functionalisme kan de structurele relaties tussen psychologische toestanden niet goed uitdrukken. De verzameling logische toestanden in een machinetabel is een eindige lijst. De verzameling mentale toestanden van personen is productief Neem alleen al de propositionele attitude 'menen dat P'. Voor P kan van alles worden ingevuld, in principe oneindig veel. Bovendien hebben bepaalde psychologische toestanden structurele overeenkomsten: 'menen dat P' heeft iets te maken met 'menen dat P en Q'. Sommige psychologische toestanden hangen sterk af van leren en geheugen. Logische toestanden hebben niet zulke structurele overeenkomsten en

zijn voor wat hun identiteit betreft niet afhankelijk van leren en geheugen (Fodor en Block, Putnam)

Al deze bezwaren geven aan dat het niet mogelijk is dat psychologische toestanden een-een corresponderen met logische toestanden, laat staan dat ze identiek zijn met logische toestanden

Fodor en Block maken een onderscheid tussen de logische toestanden van een automaat en de computationele toestanden Met deze laatste wordt bedoeld iedere toestand van de machine die uitgedrukt kan worden in termen van input, output en logische toestand (zie ook Fodor's voorzichtige definitie van functionele toestand die ik hierboven vermeldde). In dit gebruik duiden de predikaten 'heeft juist een berekening gemaakt waarin driehonderdtweeenzeventig logische toestanden zijn doorlopen', of 'heeft juist Fermat s laatste stelling bewezen', of 'heeft juist het n-de symbool van het output-alfabet getypt' allemaal mogelijke computationele toestanden van machines aan Volgens Fodor en Block geven de argumenten aan dat psychologische toestanden geen logische toestanden zijn Ze geven niet aan, met uitzondering van argument 4 dat tegen elke vorm van functionalisme pleit, dat psychologische toestanden geen computationele toestanden kunnen zijn Organismen kunnen dan nog steeds Turingmachines zijn, en hun psychologische toestanden zijn gelijk aan meer abstracte toestanden van de machine Maar er valt niet meer op een eenvoudige manier te zeggen wanneer twee organismen in dezelfde psychologische toestand zijn (Fodor 1981a, 99) De stelling dat mensen Turingmachines zouden zijn verliest daardoor veel van haar aantrekkelijkheid Het type-fysicalisme was aantrekkelijk omdat het stelde dat psychologische toestanden type-identiek waren met neurofysiologische toestanden Het bleek evenwel niet houdbaar, omdat de rechterkant van de bruguitspraken niet een neurofysiologische toestand, maar een mogelijk open disjunctie van neurofysiologische toestanden noemde Het Turingmachine-functionalisme leek opnieuw een aantrekkelijke oplossing te bieden psychologische toestanden waren type-identiek met logische toestanden Dit was compatibel met het token-fysicalisme, omdat logische toestanden zelf niet type-identiek zijn met fysische machinetoestanden Maar ook het Turingmachine-functionalisme is niet op die manier houdbaar· psychologische toestanden zijn ook niet type-identiek met logische toestanden van een Turingmachine Het blijkt

immers opnieuw dat bruguitspraken tussen psychologische toestanden en logische toestanden aan de rechterkant niet een logische toestand noemen maar een disjunctie.

Ook Putnam laat zien dat psychologische toestanden gelijk kunnen zijn aan meer abstracte toestanden van machines, bijvoorbeeld aan disjuncties, waarvan de individuele disjuncten weer bestaan uit een conjunctie van een logische toestand en een machineband. Maar hij vindt dat geen hoopvolle uitkomst. Niet alleen, meent hij, zou zo'n beschrijving oneindig zijn, maar de theorie heeft nu geen inhoud meer. Het oorspronkelijke doel was om de machinetabel te gebruiken als model voor een psychologische theorie, nu is evenwel duidelijk dat de machinetabel-beschrijving, ofschoon verschillend van een fysische beschrijving, net zo ver afstaat van de beschrijving van psychologische toestanden als een fysische beschrijving (Putnam 1973).

Het Turingmachine model kan dus niet dienen om de notie van functionele rol verder te systematiseren. Een mentale of psychologische toestand mag dan gedefinieerd zijn als die toestand die een bepaalde functionele rol speelt tussen input, output en andere mentale toestanden, ze kan noch geïdentificeerd worden (in de zin van type-identiteit) met een fysische toestand, noch met een logische toestand. De mentale soorten, gedefinieerd door hun functionele rol, vormen een dwarsdoorsnede van de soorten van zowel de fysica en de fysiologie, als van de soorten van de mathematische theorie van automaten en Turingmachines. In die zin is het functionalisme een waarlijk ontologisch neutrale theorie over het mentale, en geen fysicalistische theorie. Maar samen met het token-fysicalisme is het natuurlijk wel een fysicalistische theorie: men stelt een token-identiteit tussen iedere mentale toestand en één of andere fysische toestand Vandaar de eis dat het voor iedere functionele toestand aannemelijk moet zijn dat er een of ander mechanisme gebouwd kan worden dat die functie kan uitvoeren Alleen, als het token-fysicalisme waar is is het *altijd* aannemelijk dat zo'n mechanisme in principe gebouwd kan worden; immers, *iedere* mentale toestand is een fysische toestand en *iedere* functionele rol die zo'n toestand speelt wordt gespeeld door een fysische toestand volgens het token-fysicalisme, en iedere functie kan dus door een fysisch systeem vervuld worden. Zoals Fodor dan ook opmerkt:

> "In practice the argument usually goes in the opposite direction, if the postulation of a mental operation is essential to some cherished psychological explanation, the theorist tends to assume that there must be a program for a Turing machine that will carry out that operation" (Fodor 1981b, 130).

Zo gesteld is de eis dat een mechanisme mogelijk moet zijn voor iedere mentale toestand en mentaal proces te zwak om te laten zien dat mentale toestanden token-identiek zijn met fysische toestanden Het *geloof* in het token-fysicalisme, en niet de praktijk van het programmeren, moet dan kennelijk garanderen dat de functionalistische theorie ook een fysicalistische theorie is. Maar als men eist dat voor ieder functioneel gedefinieerd theoretisch construct een mechanisme *bestaat* dat die functie uitvoert, dan is die eis veel te sterk Er zijn, zoals we in hoofdstuk 2 zagen, heel veel cognitieve toestanden en processen die een computer niet heeft of kan uitvoeren, en die toch in de cognitieve psychologie en de filosofie van het mentale als theoretisch construct voorkomen. Nogmaals, het bestaan van de huidige Turingmachines of geprogrammeerde computers vormt geen steun voor het functionalisme *cum* token-fysicalisme (ik zal voortaan met 'functionalisme' 'functionalisme *cum* token-fysicalisme' bedoelen)

3.4.2. Functionalisme en qualia.

Volgens het functionalisme zijn mentale toestanden en processen functioneel gedefinieerde toestanden en processen. Een functionalistische definitie beperkt zich evenwel niet tot mentale toestanden en processen. Pastamachines en puntenslijpers, muizenvallen en ministers van financiën zijn allemaal in zekere zin concepten die functioneel gedefinieerd zijn, maar geen ervan is een mentaal concept zoals pijn, mening en verlangen. Kan een functionalistische definitie wel het specifieke van mentale toestanden en processen aangeven?

Het vervelende is dat er niet een kenmerk van het mentale is. We hebben in hoofdstuk 1 gezien dat Brentano meende dat intentionaliteit

het kenmerk van het psychische (of mentale) is, maar dat sommige toestanden en processen die we mentaal noemen helemaal niet intentioneel zijn. We hebben toen mentale toestanden en processen verdeeld in twee soorten: enerzijds de sensaties, die gekenmerkt worden door een kwalitatieve inhoud, en anderzijds de toestanden en processen die gekenmerkt worden door intentionaliteit. Het functionalisme komt niet met beide soorten mentale toestand even goed uit de voeten de theorie heeft problemen met de kwalitatieve aard van sensaties, zoals geïllustreerd kan worden aan het probleem van het omgekeerde spectrum Het omgekeerde spectrum, of de omgekeerde qualia, is een voorbeeld dat al in de geschriften van Locke voorkomt. Het gaat over iemand die rood ziet als groen en groen als rood, en zo door het hele kleurenspectrum, alsof hij een kleurennegatief ziet in plaats van een kleurenpositief Maar wie merkt dat ooit? Hij noemt het gras groen, net als wij, en een rijpe tomaat rood. Hij noch wij merken dat er een verschil is. Maar het verschil bestaat wel degelijk.

Het is natuurlijk mogelijk om hier een *verificationistisch* antwoord op te geven, dat wil zeggen, te antwoorden dat dit voorbeeld onzinnig is omdat het logisch onmogelijk is het bestaan van dat verschil te verifieren. Maar Putnam heeft een variant bedacht die geen problemen oplevert voor verificatie (Putnam 1981, 80). Stel je wordt op een morgen wakker en er staat een stralende blauwe zon in een strak-gele hemel, je hibiscus heeft net zijn groene bloemen ontvouwd tussen zijn rode bladeren en in de spiegel ziet je gezicht er afschuwelijk uit. Vreselijk! Maar je leert ermee te leven, je past je taalgebruik aan, langzamerhand ook de associaties die je bij de kleuren hebt, en niemand merkt er meer iets van. Alleen 's nachts in bed, huil je nog wel eens om die verloren wereld toen de kleuren nog zoals vroeger waren Je weet dat er een verandering was. Je subjectieve ervaring van rood speelt nu dezelfde functionele rol als je vroegere subjectieve ervaring van groen. De functionele rol is identiek, maar de kwalitatieve aard van de sensatie is veranderd. De functionalist kan nog zeggen dat de functionele rol niet echt hetzelfde is, vanwege je sporadische nachtelijke droefenis. Maar stel je dan voor dat je na je aanpassing aan je nieuwe sensaties een aanval van amnesie krijgt die elke herinnering aan hoe de kleuren vroeger waren uitwist. Dan valt niet meer te ontkennen dat de sensatie die je nù een 'sensatie van

blauw' noemt precies dezelfde functionele rol speelt als de sensatie die je vroeger een 'sensatie van blauw' noemde vroeger speelde, terwijl hij een totaal verschillende aard heeft De kwaliteit is veranderd De kwaliteit is hier geen functionele toestand, en valt niet te vangen in termen van het functionalisme

Er is nog een andere verdedigingslijn open voor de functionalist. Ze zou kunnen zeggen dat we twee functioneel identieke mentale toestanden gewoon als identiek beschouwen, ondanks eventuele kwalitatieve verschillen. Want verschillen in kwalitatieve inhoud van mentale toestanden die met geen enkel functioneel verschil corresponderen zijn *ipso facto* niet relevant voor theorieconstructie in de psychologie, en spelen verder geen enkele rol. Stel je voor dat het zo is dat iedereen in feite verschillende qualia heeft in de functioneel gedefinieerde toestand van bijvoorbeeld pijn. In dat geval zou het redelijk zijn te zeggen dat de aard van de qualia van een organisme irrelevant is voor de kwestie of het pijn heeft Maar dit argument kan tot vervelende consequenties leiden Het kan best mogelijk zijn dat twee mentale toestanden functioneel identiek zijn (dat wil zeggen, dezelfde relaties hebben met input, output en andere mentale toestanden), zelfs als slechts één van de twee toestanden een kwalitatieve inhoud heeft In dat geval zou je moeten zeggen dat een organisme pijn heeft hoewel het *helemaal* niets voelt En dat is een consequentie die wel volkomen onacceptabel lijkt (Fodor and Block 1972, herdrukt in Fodor 1981a).

Het voorbeeld van het omgekeerde spectrum en zijn varianten zijn meer dan een verbaal grapje Kwalitatieve inhoud is het kenmerk van een belangrijke groep van mentale toestanden en processen, en is vaak het belangrijkste punt geweest in de discussies rond het lichaam-geest probleem In de discussies rond de identiteitstheorie ging het juist om *qualia,* om sensaties, om zogenaamde *raw feels.* Men vroeg zich af of trachtte te verduidelijken hoe het mogelijk was dat een bepaalde hersentoestand gevoeld werd als een sensatie. Het functionalisme heeft evenwel nauwelijks iets of helemaal niets te zeggen over sensaties. En al is het probleem van de kwalitatieve inhoud wellicht voor psychologen van ondergeschikt belang, het "poses a serious threat to the assertion that functonalism can provide a general theory of the mental" (Fodor 1981b, 130). Toch wordt het probleem niet vaak gethematiseerd, men

stelt het zoeken naar een oplossing uit (b v Fodor 1981b), of zoekt deze in de richting van een aanvulling van het functionalisme met wat identiteitstheorie (b.v Putnam 1981) of met een eliminatief materialisme (b.v Dennett 1978a, 1978b (35), zie ook over het qualiaprobleem Gunderson 1971, Nagel 1974, Block 1978, Shoemaker 1979, Armstrong & Malcolm 1984) Functionalisten concentreren zich veel meer op de mentale toestanden die gekenmerkt zijn door intentionaliteit Op dat gebied vindt men volgens velen de belangrijke successen van de cognitiewetenschap Intentionaliteit, zo heb ik boven al gezegd, heeft, op zijn ruimst gezegd, iets te maken met de gerichtheid op iets ik kan niet geloven zonder *iets* te geloven enz. Het functionalisme weet misschien geen raad met qualia, zoals een pure sensatie van pijn, of van blauw, of van angst, maar het meent wel veel te zeggen te hebben over mentale toestanden met intentionaliteit, zoals 'geloven of menen dat P' (36) en 'willen dat P' Het functionalisme heeft zich geconcentreerd op de mentale toestanden en processen die gekenmerkt worden door intentionaliteit, en sluit daarbij aan bij de zogenaamde *belief-desire* psychologie, de psychologie die het gedrag van een systeem wil verklaren met een beroep op wat het systeem gelooft of meent en wil De huidige cognitieve psychologie vindt, in tegenstelling tot het behaviorisme, dat zulke verklaringen legitiem zijn Theorieen over leren en perceptie, bijvoorbeeld, gaan voornamelijk over de vraag hoe allerlei meningen die een organisme heeft bepaald worden door de aard van zijn ervaringen en van zijn genetisch bepaalde mogelijkheden

De functionalistische, cognitieve theorie van het mentale wil verklaren wat er gebeurt in een geval waarvan men een alledaagse verklaringen kan geven als "Hamlet vermoordde de man achter het scherm omdat hij meende dat het zijn oom was" Ze vindt het zinvol om mentale termen te gebruiken in de psychologie voor functioneel gedefinieerde toestanden en processen, ze wil de notie van mentale veroorzaking onderzoeken, en daarbij streeft ze ernaar toch een fysicalistische theorie van het mentale te zijn.

Het functionalisme, aldus gekarakteriseerd, kan op twee manieren verder uitgewerkt worden Enerzijds kan men beweren dat die functioneel gedefinieerde mentale toestanden en processen echt bestaan in een organisme, anderzijds kan men volhouden dat mentale toestanden

en processen niet echt bestaan in een organisme, maar er om praktische redenen aan toegeschreven kunnen worden. (We hebben in 1.3.3.2 gezien dat Boden beide mogelijkheden open hield.) De realistische variant van het functionalisme wordt in het volgende hoofdstuk behandeld aan de hand van een analyse van het werk van Fodor, en de instrumentalistische variant wordt in hoofdstuk 5 uitgewerkt aan de hand van een analyse van het werk van Dennett.

4. JERRY FODOR "I'M THE ONLY PRESIDENT YOU'VE GOT".

4 1. Inleiding.

De filosoof-psycholoog-linguïst Jerry Fodor heeft geprobeerd een theorie van het mentale te ontwikkelen die de cognitieve psychologie tot grondslag kan dienen. Hij baseert zich daarbij op de algemene uitgangspunten van het functionalisme en het token-fysicalisme, zoals die in hoofdstuk 3 zijn uiteengezet Al in zijn boek *Psychological explanation* uit 1968 laat hij zien dat het toelaten van mentale termen in de psychologie niet hoeft te leiden tot dualisme (contra het behaviorisme), en onderzoekt hij de mogelijkheid voor causale verklaringen in de psychologie en voor een functionele definitie van mentale toestanden en processen In zijn boek *The language of thought* (1975) ontvouwt hij zijn theorie van het mentale, en in latere artikelen (deels gebundeld in Fodor 1981a) werkt hij haar in verschillende richtingen uit Hij probeert de theorie van het mentale waar de cognitieve psychologie ons op vast legt te expliciteren (Fodor 1975, IX), waarbij hij niet terugschrikt voor op het eerste gezicht "wilde" ideeen· "Remotely plausible theories are better than no theories at all" (Fodor 1975, 27).

Fodor is een realist voor wat betreft psychologische of propositionele attitudes zoals menen en wensen enz Mensen hebben echt mentale toestanden die gekenmerkt worden door intentionaliteit, en het is niet alleen maar handig om ze zulke toestanden instrumentalistisch toe te schrijven. Daaruit volgt, volgens Fodor, dat mensen interne representaties moeten hebben, die ook echt bestaan. In 4.2 zet ik Fodor's theorie over propositionele attitudes en interne representaties uiteen

Mentale toestanden en processen zijn functioneel gedefinieerd, en spelen volgens Fodor een causale rol in relatie tot gedrag. Dit houdt in dat de interne representaties een causale rol moeten kunnen spelen. In 4.3 wil ik laten zien hoe dit volgens Fodor mogelijk is: hij moet dan zijn representationele theorie van het mentale preciseren tot een computationele theorie van het mentale Deze theorie van het mentale dicteert dat mentale toestanden in intentionele termen beschreven

moeten worden Het gaat er in Fodor's theorie om hoe het te verklaren systeem (mens) de wereld voor zichzelf representeert, dat is wat het gedrag (mede)veroorzaakt, en niet de wereld zelf In 4 4 zet ik uiteen wat dit dictaat voor de beschrijving van mentale toestanden volgens Fodor inhoudt voor de methodologie en de mogelijkheid van verschillende soorten psychologie

Tot zover zal ik niet veel of diepgaande kritiek hebben op Fodor's argumenten Er valt wel iets voor te zeggen dat Fodor s theorie van het mentale de enige mogelijke is voor een cognitieve psychologie. Maar de theorie, gepresenteerd als een *fysicalistische* theorie van het mentale, die een *fysicalistische* oplossing moet kunnen bieden voor het lichaam-geest probleem, laat drie enorme problemen onopgelost
Ten eerste. wat maakt interne representaties tot representaties *van de buitenwereld?* Ik noem dit het referentieprobleem.
Ten tweede· wat maakt dat interne representaties überhaupt *iets representeren,* inhoud hebben? Ik noem dit het betekenisprobleem (37).
En ten derde: *voor wie* representeren de interne representaties iets? Ik noem dit het intentionaliteitsprobleem.

In 4.5 zal ik deze drie problemen bespreken en laten zien dat Fodor de eerste twee onvoldoende uit elkaar houdt en het derde (en grootste) probleem veronachtzaamt. In 4.6 bespreek en bekritiseer ik Fodor's recente pogingen om de eerste twee pogingen op te lossen. In 4.7 zal ik concluderen dat Fodor's theorie geen fysicalistische oplossing biedt voor het lichaam-geest probleem.

4.2. Propositionele attitudes en interne representaties.

Fodor's theorie concentreert zich, zoals het functionalisme en de cognitiewetenschap in het algemeen (zie 1.3.2 en 3 4.2), op de mentale toestanden die gekenmerkt worden door intentionaliteit, zoals menen en wensen enz. Hij is een realist voor wat betreft die toestanden· de toestand van menen dat p kan volgens hem aan mensen letterlijk, realistisch, worden toegeschreven. Zijn argumenten voor dit realisme zijn volgens mij even eenvoudig als steekhoudend

Allereerst wil hij afzien van *algemene* argumenten voor het

wetenschappelijk realisme, omdat juist het feit dat ze algemeen zijn ze irrelevent maakt voor de discussie over de vraag of juist *mentale* toestanden en processen al dan niet bestaan (Fodor 1981a, 115) En vervolgens merkt hij op dat het wel heel vreemd zou zijn als er geen meningen en wensen zouden bestaan. Alle (niet-behavioristische) psychologische theorieen noemen ze in hun verklaringen, kinderen schrijven ze toe zonder dat expliciet te hoeven leren, en niemand twijfelt *serieus* (in tegenstelling tot filosofisch) aan hun bestaan (1981a, 122).

> " . Why on earth should one *not* have the belief when all the evidence conspires to tempt one to it? The burden of argument lies, surely, upon those who say that it is not the case that one believes even what one is strongly disposed to believe that one believes" (Fodor 1981a, 102).

Dit is een argument van het type· "Geloof je dat zelf nou echt?", en ofschoon het geen *knock down* argument is, vind ik het een uitstekende manier om de bewijslast te leggen waar hij hoort. bij de verdedigers van een tegen-intuitieve theorie (in hoofdstuk 5 komt zo'n verdediger, Dennett, uitgebreid aan het woord)

Aangezien Fodor aanneemt dat mentale toestanden zoals menen en verwachten echt bestaan, zal hij moeten aangeven wat dat voor toestanden zijn Wanneer bijvoorbeeld Hamlet meent dat zijn oom achter het scherm staat, dan heeft dat te maken met een soort relatie van menen tussen Hamlet en een propositie die de inhoud van dat menen vormt (namelijk de propositie dat zijn oom achter het scherm staat) Zulke mentale toestanden als deze worden 'propositionele attitudes' genoemd, een term van Russell (1940). Er zijn drie vrijheidsgraden in de formulering van een propositionele attitude persoon, attitude-type en propositie, *x* meent dat p of *y* meent dat p; x *meent* dat p of *vreest* dat p of *hoopt* dat p; x meent dat *p* of dat *q* enz.

Er zijn andere manieren om naar propositionele attitudes te verwijzen, zoals 'de mening die Marie deed blozen' of 'Lubbers' meest controversiële mening', maar die manieren zijn parasitair; men kan na zo'n verwijzing doorgaan en vragen: "En welke mening is dat?" in de hoop een *identificerende* verwijzing te krijgen - bijvoorbeeld 'Marie's

mening dat Jan haar geheim kende' Sommige mensen denken dat we ooit in staat zullen zijn om meningen neurofysiologisch te identificeren ('de mening in Jan's cortex met fysisch kenmerk F'), maar op het ogenblik hebben we in ieder geval geen manier om een mening aan te duiden op de manier waarop we een boek aan kunnen duiden door een fysische beschrijving van één van de exemplaren ervan te geven Een propositionele attitude moet worden aangeduid door persoon, attitude-type en propositie te vermelden.

In een artikel getiteld 'Propositional attitudes' (in Fodor 1981a) zet Fodor zijn theorie over propositionele attitudes het beknoptst uiteen. (Uitgebreider is deze theorie te vinden in Fodor 1975.) Hij somt de voorwaarden op waaraan een theorie van propositionele attitudes moet voldoen. Vervolgens laat hij zien dat de theorie van propositionele attitudes het bestaan van interne representaties nodig heeft. Ik geef in het volgende zijn redenering weer:

Propositionele attitudes moeten geanalyseerd worden als *relaties*. Het werkwoord 'menen' in zinnen als 'Jan meent dat het regent' drukt een relatie uit tussen Jan en iets anders Wanneer men aanneemt dat een werkwoord van propositionele attitude geen twee-plaatsige relatie is maar een een-plaatsig predikaat, waarbij het werkwoord semantisch 'gefuseerd' is met het lijdend voorwerp, dan gaan allerlei verbanden verloren. Zo'n fusie-opvatting (zie b v. Loar 1981) kan geen verband zien tussen 'menen-dat-p' en 'menen-dat-q', of 'menen-dat-p' en 'vrezen-dat-p'; volgens die opvatting gaat het hier om drie verschillende een-plaatsige predikaten zonder interne structuur. 'Jan meent dat het regent' heeft volgens de fusie-opvatting net zo weinig te maken met 'het regent' als 'olie' met 'katholiek' (zie b.v. Loar 1981, zie ook noot 13). Het zou dan ook toeval zijn dat 'Jan meent dat het regent' inhoudt dat Jan iets waars meent als 'Het regent' waar is

Propositionele attitudes zijn dus relaties tussen personen en iets anders. Maar tussen personen en wat? Het meest voor de hand liggend zijn *proposities*. Proposities kunnen echter niet de onmiddellijke objecten van propositionele attitudes zijn, volgens Fodor, om twee redenen Ten eerste, proposities hebben niet de juiste eigenschappen, met name hebben ze geen vorm Proposities neutraliseren de lexico-syntactische (vorm-)verschillen tussen verschillende manieren om hetzelfde te zeggen. Ten tweede: wat betekent het te zeggen:

"Propositionele attitudes zijn relaties tot proposities"? Hoe kun je in een relatie tot een propositie staan, hoe kan de geest een propositie vatten?

Als volgende mogelijkheid oppert Fodor de theorie van Carnap Deze heeft in *Meaning and necessity* (1947) gesuggereerd dat propositionele attitudes geconstrueerd kunnen worden als relaties tussen mensen en *zinnen die ze gedisponeerd zijn te uiten;* bijvoorbeeld, tussen mensen en zinnen uit het Nederlands (of Engels). Deze theorie laat goed uitkomen dat propositionele attitudes relaties zijn, en, aangezien zinnen een interne structuur hebben, dat verschillende attitudes inhoudelijke verbanden kunnen vertonen Bovendien laat de theorie zien waarom toeschrijvingen van propositionele attitudes intensioneel (in de zin van Chisholm) zijn Als Jan iets zegt, zegt hij een bepaalde zin, b.v. "Marie is ziek" en niet een andere, daarmee vergelijkbare zin zoals "Iemand is ziek". Hetzelfde geldt dan als Jan gedisponeerd is om een bepaalde zin te zeggen.

Ondanks deze successen valt er toch nog wel het een en ander aan te merken op Carnap's theorie. Volgens Carnap hebben twee mensen dezelfde mening dan en slechts dan als ze gedisponeerd zijn dezelfde zin te zeggen. Dit levert soms een te fijne onderscheiding op. Uit allerlei sorteertaken en conceptvormingstaken blijkt bijvoorbeeld dat de mentale toestand van menen dat P en Q eenvoudiger is dan de mentale toestand van menen dat noch niet-P noch niet-Q. Wanneer menen een relatie met een bepaalde zin (en niet met een bepaalde propositie) inhoudt, dan maakt dit inderdaad verschil, ook al hebben '$P \wedge Q$' en '$\sim(\sim P \vee \sim Q)$' dezelfde waarheidswaarde (Wason & Johnson-Laird, 1972). Maar zou je ook willen zeggen dat de zin 'Jan meent dat Marie Wim gebeten heeft' een andere mening toeschrijft dan 'Jan meent dat Wim door Marie gebeten is'?

Voorts kunnen volgens deze theorie anderstaligen niet dezelfde mening hebben, en pretalige kinderen en dieren helemaal geen meningen hebben. Men mag toch aannemen dat ook poezen en pretalige kinderen iets dergelijks kunnen als menen dat het regent. Ooit een poes in de regen gezien? Tenslotte, als zinnen van een natuurlijke taal het object zijn van propositionele attitudes, hoe wordt dan de eerste taal geleerd? Iedere niet-behavioristische theorie op dit gebied beschrijft het leren van een eerste taal in termen van verwachtingen

en hypothese toetsen enz En dat betekent dat er tenminste sommige propositionele attitudes moeten zijn die niet een relatie inhouden met zinnen van een natuurlijke taal, want de nieuwe taal-leerder heeft die zinnen nog niet tot zijn beschikking (dit punt wordt zeer uitvoerig behandeld in Fodor 1975)

Alles bij elkaar lijkt de situatie ontmoedigend Enerzijds zijn er argumenten om Carnap's theorie te accepteren, anderzijds zijn er even plausibele argumenten om de theorie niet te accepteren Maar Fodor ziet een goede mogelijkheid om de theorie gedeeltelijk te behouden, en slechts voor een deel te verwerpen Wat hij wil behouden is de opvatting dat de objecten van propositionele attitudes *zinnen* zijn, met een logische vorm, waarheidswaarden enz Wat hij wil verwerpen, omdat dat volgens hem de problemen oplevert, is dat de objecten van propositionele attitudes zinnen *van een natuurlijke taal* zijn Zijn voorstel is om de objecten van propositionele attitudes zinnen te laten zijn van een niet-natuurlijke taal, zinnen van wat hij in zijn boek uit 1975 the language of thought noemt, een Intern Representatie Systeem of ook wel het Mentalees

Propositionele attitudes zijn nu relaties met interne representaties, het zijn dus relaties, en toeschrijvingen ervan zijn intensioneel omdat propositionele attitudes relaties met bepaalde representaties zijn Wanneer Nederlands-taligen menen dat het regent, zijn ze in relatie met een teken van een formule in het Mentalees, en als ze willen zeggen wat ze dan menen, zeggen ze 'Het regent'. Dezelfde interne formule, noem hem F(het regent), ligt ten grondslag aan zowel het menen als het zeggen Zo wordt een natuurlijke taal gezien als een conventioneel systeem voor het uitdrukken van gedachten Die gedachtengangen zelf verlopen in een niet-natuurlijke taal

Neem nu eens aan dat de interne taal aangeboren is, dat de formules ervan een-op-een corresponderen met de objecten van propositionele attitudes, en dat hij universeel is Dan gaan de bezwaren die tegen Carnaps theorie waren in te brengen niet meer op 'Marie beet Wim' en 'Wim werd door Marie gebeten' corresponderen nu met dezelfde interne zin, maar 'P∧Q' en '~(~P∨~Q)' corresponderen met verschillende zinnen Iedereen die onze mentale toestanden en processen deelt, deelt ook ons systeem van interne representaties

Dit is Fodor's theorie van propositionele attitudes propositionele

attitudes zijn relaties tot zinnen in een intern representatiesysteem. Het zijn gemedieerde relaties tot proposities, gemedieerd via de interne representaties, die die proposities uitdrukken.

> "As for me", zegt Fodor, "I want a *mechanism* for the relation between organisms and propositions, and the only one I can think of is mediation by internal representations" (Fodor 1981a, 202).

4.3. Mentale veroorzaking en formaliteit.

Volgens Fodor kunnen propositionele attitudes realistisch worden toegeschreven aan mensen. mensen hebben echt meningen en wensen en gedachten. En een propositionele attitude is een relatie tot een interne representatie, dus mensen hebben ook echt interne representaties. Volgens Fodor kan een cognitieve psychologie er niet onderuit om deze *representationele theorie van het mentale* aan te hangen. Zo begint hij zijn artikel 'Methodological solipsism considered as a research strategy in cognitive psychology' uit 1980 met de volgende zinnen:

> "Your standard contemporary cognitive psychologist - your thoroughly modern mentalist - is disposed to reason as follows. To think (e.g) that Marvin is melancholy is to represent Marvin in a certain way, viz., as being melancholy (and not, for example, as being maudlin, morose, moody or merely moping and dyspeptic). But surely we cannot represent Marvin as being melancholy except as we are in some or other relation to a representation of Marvin; and not just to *any* representation of Marvin, but, in particular, to a representation the content of which is *that* Marvin is melancholy; a representation which, as it were, expresses the proposition that Marvin is melancholy" (Fodor 1980a, 63)

De reëel bestaande interne representaties (Fodor lijkt te denken aan

een neurale code (b.v. 1978)) vormen een mechanisme voor de relatie tussen organismen en proposities, proposities zijn volgens hem ongrijpbare (en onbegrijpelijke) entiteiten, maar interne representaties zijn concreet genoeg. Er is nog een andere reden waarom het volgens Fodor goed is dat er echte interne representaties bestaan. Hij wil laten zien dat mentale toestanden en processen een rol spelen in de veroorzaking van gedrag: "Hamlet vermoordde de man achter het scherm omdat hij meende dat het zijn oom was (en omdat hij meende dat zijn oom zijn vader vermoord had, en omdat hij zijn vader wenste te wreken enz.)" Mentale *toestanden*, althans propositionele attitudes, kunnen gezien worden als relaties tot interne representaties. Welnu, mentale *processen* kunnen gezien worden als operaties die gedefinieerd worden over die representaties Die processen zijn symboolmanipulaties, en die processen veroorzaken (mede) het gedrag. Ook hiervoor kunnen we een mechanisme postuleren, wanneer we zeggen dat mentale processen *computationeel* zijn. Computationele processen kennen we immers uit de computerwereld. er bestaan mechanismen die computationele processen uitvoeren

Volgens Fodor moet de cognitieve psychologie haar representationele theorie van het mentale aanvullen met (of misschien preciseren tot) een *computationele theorie van het mentale.* De representationele theorie laat zien wat het is om in een propositionele attitude te zijn, en hoe een organisme in relatie kan staan tot een propositie. Gecombineerd met de computationele theorie laat ze zien hoe mentale toestanden en processen een rol kunnen spelen in de veroorzaking van gedrag. Fodor's argument om de representationele theorie van het mentale te preciseren tot een computationele theorie, is dat hij wil verklaren hoe mentale veroorzaking mogelijk is. En net als voor de relatie tussen organismen en proposities (zie 4 2) wil hij een *mechanisme* om mentale veroorzaking te verklaren. Het enige mechanisme dat hij kent dat mentale veroorzaking zou kunnen verklaren is een computationeel systeem En daarom kiest hij voor een computationele theorie van het mentale.

Volgens de representationele theorie van het mentale kunnen meningen met verschillende inhoud verschillende effecten hebben op het gedrag De computationele theorie kan als volgt verklaren hoe dat kan· mentale processen zijn computationeel, dat wil zeggen, mentale

processen zijn zowel *symbolisch* als *formeel*. Ze zijn symbolisch omdat ze gedefinieerd zijn over symbolen: de interne representaties. En ze zijn formeel omdat ze werken op de *vorm* van de objecten waarover ze gedefinieerd zijn, ze kunnen die objecten alleen op hun vorm herkennen. Formele operaties zijn van toepassing op grond van de vorm van hun symbolische objecten, ongeacht de inhoud van die symbolen, zulke operaties hebben geen toegang tot die inhoud

> "Formal operations are the ones that are specified without reference to such semantic properties of representations as, for example, truth, reference and meaning ... formal operations apply in terms of the, as it were, "shapes" of the objects in their domains" (Fodor 1980a, 64).

Computationele operaties zijn in deze zin formeel; bijvoorbeeld in een computerprogramma zijn instructies alleen verschillend als ze minstens een gat in de ponskaarten niet delen. Als nu de computationele theorie van het mentale waar is, dan kunnen de mentale processen, die computationeel en dus formeel zijn, inhoudsverschillen van de mentale representaties niet in acht nemen. Maar hoe kunnen dan meningen met verschillende inhoud verschillende effecten hebben op het gedrag? Dat kan alleen als in de interne representaties inhoudsverschillen corresponderen met vormverschillen De computationele theorie van het mentale vereist dat twee gedachten alleen verschillen in inhoud als ze geïdentificeerd worden met relaties tot *formeel* verschillende representaties. Wanneer subject en relatie-type gelijk zijn, dan kunnen mentale toestanden alleen verschillen als de representaties die hun object vormen formeel verschillend zijn, een verschillende vorm hebben.

Fodor spreekt van een formaliteitsconditie ten aanzien van mentale toestanden: mentale toestanden moeten ingedeeld worden op grond van hun formele kenmerken. In zijn Solipsisme-artikel (1980a) wil hij laten zien dat de psychologie op onafhankelijke gronden de formaliteitsconditie moet honoreren en dat de computationele theorie van het mentale, die de formaliteitsconditie afdwingt, wat dat betreft goed zit.

Fodor merkt in een ander artikel ergens op: "... it is frightfully

easy to find oneself arguing backwards" (Fodor 1981a, 112), en *arguing backwards* is precies wat hij hier doet Want wat hij in feite laat zien is dat de psychologie een *opake* - niet een formele - indeling van mentale toestanden nodig heeft, en dat is een veel zwakkere conditie dan de formaliteitsconditie De enige reden om een formaliteitsconditie aan te hangen is dat de computationele theorie je daartoe dwingt, omdat computers hun symbolen manipuleren op grond van hun vorm. En het enige argument dat Fodor geeft voor een computationele theorie is dat hij zich mentale veroorzaking niet anders kan voorstellen.

Fodor's argumenten voor een *opake* taxonomie van mentale toestanden zijn sterk. Al in 1.3.2 hebben we gezien dat alledaagse toeschrijvingen van mentale toestanden voldoen aan Chisholm's criteria voor intensionele zinnen Fodor definieert een opake context als een context waarin existentiele generalisatie en/of substitutie van coreferentiele termen niet opgaat. Volgens een opake taxonomie zijn 'ik zal de man achter het scherm doden' en 'ik zal Polonius doden' geen instanties van hetzelfde voornemen, ook al is de man achter het scherm Polonius. Macbeth's mening 'daar is een dolk' als hij een dolk ziet en zijn mening 'daar is een dolk' als hij een dolk hallucineert zijn wel instanties van dezelfde mening.

Maar de psychologie heeft niet alleen een opake taxonomie van mentale toestanden nodig omdat men in het dagelijks leven mentale toestanden zo indeelt. Een opake taxonomie is ook nodig voor een theorie van mentale veroorzaking van gedrag. Want hoe de buitenwereld in feite is maakt voor de mentale veroorzaking van gedrag niet uit. Hamlet vermoordde de man achter het scherm omdat hij meende dat het zijn oom was. En Lady Macbeth wrong haar handen omdat ze meende dat er bloed aan kleefde. In werkelijkheid was het Polonius achter het scherm en waren de Lady's handen, althans letterlijk, lelieblank. Maar de onwaarheid van hun meningen maakte voor hun gedrag geen verschil Let wel, zo'n indeling van mentale toestanden is opaak vanuit het standpunt van de buitenstaander, de beschrijver van de mentale toestanden. Voor degene die de mentale toestanden heeft bestaat het onderscheid tussen transparant en opaak niet. Als *ik* een dolk meen te zien dan maak ik wel de existentiële generalisatie dat er een dolk is, mijn wens om met mijn geliefde te

trouwen was transparant voor alle beschrijvingen van hem die, voor zover ik wist, coreferentieel waren Alleen voor een ander die mijn mentale toestanden beschrijft zijn existentiele generalisatie en substitutie van coreferentiele termen niet toegestaan De onderzoeker moet een opake taxonomie van mentale toestanden aanhouden om het gedrag van zijn proefpersonen causaal te kunnen verklaren. Het gaat er immers om hoe de persoon de objecten van zijn verlangens en meningen voor zichzelf representeert Zoals Fodor zegt (over weer een ander literair voorbeeld)

> " what O *does,* how he in the proprietary sense behaves, will depend on which description he (literally) had in mind If it s "Jocasta", courtship behavior follows *ceteris paribus*. Whereas, if it's "my Mum", we have the situation towards the end of the play and Oedipus at Colonus eventually ensues" (Fodor 1980a, 66).

Zo weet Fodor aannemelijk te maken dat de psychologie een opake taxonomie van mentale toestanden nodig heeft, omdat de waarheid of de referentie van interne representaties geen rol spelen in de veroorzaking van het gedrag Als nu de computationele theorie van het mentale waar is, dan hebben de mentale, computationele processen alleen toegang tot de formele eigenschappen van interne representaties. Ze hebben dus geen toegang tot de *semantische* eigenschappen van zulke representaties, inclusief de eigenschap waar te zijn, of de eigenschap uberhaupt een representatie *van iets in de wereld* te zijn. In de computationele processen spelen de interne representaties hun rol ongeacht of ze waar zijn, of dat het hallucinaties zijn Dat maakt immers voor de veroorzaking van het gedrag niet uit In zoverre is de computationele theorie compatibel met een opake taxonomie van mentale toestanden. Voor de mentale processen maakt de waarheid van een representatie geen verschil

Maar de mentale processen hebben volgens de computationele theorie helemaal geen toegang tot de semantische eigenschappen van de representaties, ook niet tot hun betekenis! Ze reageren uitsluitend op de vorm van de representaties De formaliteitsconditie is veel sterker dan de eis van een opake taxonomie van mentale toestanden, en Fodor

heeft nog nergens laten zien dat de psychologie die sterkere conditie nodig heeft.

In feite bestaat Fodor's argument voor een formaliteitsconditie slechts uit eén zin: "... to put it mildly, it's hard to see how internal representations could differ in causal role *unless* they differed in form" (Fodor 1980a, 68). Aangezien hij een mechanisme wil hebben om iets te kunnen verklaren moeten mentale oorzaken toch een soort fysische oorzaken zijn.

Fodor weet het echter te doen voorkomen alsof zijn goede argumenten voor een opake taxonomie van mentale toestanden *ipso facto* argumenten voor een formaliteitsconditie zijn, en doet heel tevreden over de combinatie van de representationele en de computationele theorie van het mentale.

Wanneer we nu de representationele theorie van het mentale en de computationele theorie van het mentale samen nemen, aldus Fodor, dan kunnen we zeggen dat mentale representaties het gedrag beinvloeden op grond van hun inhoud, waarbij we moeten stellen dat mentale representaties alleen verschillen van inhoud als ze ook verschillen in vorm. Het eerste is nodig om plausibel te maken dat mentale toestanden relaties zijn tot mentale representaties, en het tweede is nodig om plausibel te maken dat mentale processen computaties zijn Computaties zijn immers processen waarbij representaties hun gevolgen hebben enkel op grond van hun vorm. Fodor zegt dan:

> "By thus exploiting the notions of content and computation *together,* a cognitive theory seeks to connect the *intensional* properties of mental states with their *causal* properties vis-à-vis behavior. Which is, of course, exactly what a theory of the mind ought to do" (16) (Fodor 1980a, 68).

4.4. Methodologisch solipsisme en twee soorten psychologie.

In zijn 'Methodologisch solipsisme' wil Fodor laten zien hoe zijn formaliteitsconditie vorm geeft aan de cognitieve psychologie. Een psychologische theorie van de mentale veroorzaking van gedrag die

voldoet aan de formaliteitsconditie impliceert een methodologisch solipsisme (de term is van Putnam 1975): zo'n psychologie kan afzien van het bestaan van alles en iedereen behalve degene wiens mentale toestanden en processen en wiens gedrag bestudeerd worden. Als mentale processen formeel zijn, dan hebben ze alleen toegang tot de formele eigenschappen van de representaties van de omgeving die de zintuigen leveren. Ze hebben dus geen toegang tot de *semantische* eigenschappen van zulke representaties, inclusief hun waarheid of referentie.

Aan de hand van dit methodologisch solipsisme onderscheidt Fodor twee grote tradities in de geschiedenis van de psychologie: 'Rationele psychologie' aan de ene kant en 'Naturalisme' aan de andere kant. De rationele psychologie gaat terug op Descartes. Descartes beweert dat het in zekere zin voor je mentale toestanden niet uitmaakt hoe de wereld werkelijk is Hoe weet ik zeker dat ik nu schrijf, en niet alleen maar droom dat ik schrijf? Al mijn ervaringen, en, *a fortiori*, mijn meningen, zouden precies hetzelfde kunnen zijn ook al was de wereld heel anders dan hij nu is.

Aan de andere kant is er een traditie die beweert dat het een strategische vergissing is te proberen een psychologie te ontwikkelen die mentale toestanden individualiseert zonder verwijzing naar hun oorzaken en gevolgen in de omgeving. Fodor denkt daarbij aan de Amerikaanse Naturalisten, in het bijzonder Peirce en Dewey, aan de leertheoretici, en aan tijdgenoten zoals Quine in de filosofie en Gibson in de psychologie. Volgens deze traditie is de psychologie een tak van de biologie, en moet men het organisme beschouwen in zijn fysische omgeving. Het is juist de taak van de psycholoog om de organisme/omgeving-interacties te bestuderen die het gedrag uitmaken.

Dat een zeker methodologisch solipsisme geïmpliceerd is in veel cognitieve psychologie, en dat deze psychologie tot de rationele psychologie gerekend moet worden, laat Fodor zien aan de hand van een voorbeeld. Hij bespreekt het beroemde programma Shrdlu van Terry Winograd (Winograd 1971).

Shrdlu is een programma voor een computer die 'leeft in' en 'interacteert met' een eenvoudige wereld van geometrische, gekleurde blokken. De computer kan de blokken volgens opdracht arrangeren, hij kan 'perceptuele' rapporten van de huidige stand van zijn omgeving

geven en geheugen'-rapporten over vroegere toestanden, hij kan eenvoudige plannen opstellen om een bepaald blokkenarrangement tot stand te brengen, en hij kan in redelijk Engels 'converseren' over al deze zaken.

Het interessante is nu, dat de machine-omgeving die het object is van deze handelingen en conversaties in feite niet bestaat De programmeur arrangeert de geheugentoestanden van de machine zodanig dat alle gegevens zo zijn *alsof* er een blokkenwereld is De machine leeft helemaal niet in een blokkenwereld, al zijn meningen zijn onwaar Maar dat doet er voor de machine niet toe Een computer voldoet aan de formaliteitsconditie de computaties hebben alleen toegang tot de formele (niet-semantische) eigenschappen van de representaties die gemanipuleerd worden

> "In effect" zegt Fodor, "the device is in precisely the situation that Descartes dreads, it's a mere computer which dreams that it's a robot" (Fodor 1980a, 65)

En even verderop beweert hij over een computer

> "The machine has no access to that interpretation, and its computations are in no way affected by it The machine doesn't know what it's talking about, it doesn't care, *about* is a semantic relation" (Fodor 1980a, 65)

Maar natuurlijk weten mensen wel waar ze over praten, meent Fodor. De computer die print "Robin Roberts won 28" in antwoord op een vraag over honkbal verwijst niet naar RR Maar als Fodor hetzelfde zegt verwijst hij wel. Zijn mentale representatie van RR heeft wel degelijk iets met RR te maken. En dat 'iets te maken hebben' heeft van doen met bepaalde relaties tussen RR en JF in de wereld Volgens Fodor is er een *causale* keten van gebeurtenissen tussen distale stimulus en proximale representatie (Fodor 1975, 204) En hier komt dan een naturalistische psychologie in het spel, die zich bezighoudt met organisme/omgeving transacties.

Zo te zien vullen rationele en naturalistische psychologie elkaar aan· de eerste gaat over formele processen gedefinieerd over mentale

representaties, de tweede gaat over de (vermoedelijk causale) relaties tussen representaties en de wereld die de semantische interpretaties van die representaties vastleggen

Fodor heeft evenwel geen theorie over de semantische interpretaties van representaties, althans, tot 1981 zegt hij dat In de inleiding van zijn *Representations* (1981a) zegt hij zo'n semantische theorie niet te hebben, en zowel in zijn solipsisme-artikel als in een artikel getiteld 'Tom Swift and his procedural grandmother' (1978, herdrukt in 1981a) beweert hij heel expliciet dat de cognitiewetenschap en de AI *juist niet* werken aan de vraag hoe interne representaties gerelateerd zijn aan de wereld

In zijn solipsisme-artikel wil Fodor laten zien waarom zo'n semantische theorie voor interne representaties er niet is in de cognitiewetenschap. Niet alleen is de cognitieve psychologie onderworpen aan de formaliteitsconditie, en houdt ze zich dus niet bezig met de semantische eigenschappen van representaties, maar een psychologie die zich wel met die semantische eigenschappen wil bezighouden is volgens hem praktisch onmogelijk, althans voorlopig. Fodor schetst hoe een theorie over de relaties tussen representaties en de buitenwereld er moet uitzien volgens hem·

Wat mijn gedachte over Robbie Robertson maakt tot een gedachte *over* Robbie Robertson (Fodor kent een Robin Roberts, baseball pitcher, maar ik ken, met die initialen, alleen een Robbie Robertson, zanger/gitarist, of Ronald Reagan, acteur/president), is een of andere causale relatie tussen hem en mij. We hebben een beschrijving van RR nodig zodat de causale relatie opgaat op grond van het feit dat RR voldoet aan die beschrijving We hebben een beschrijving F nodig zodat geldt 'het feit dat x F is verklaart causaal dat y een representatie is van x'. En wat zou zo'n beschrijving kunnen zijn?

In sommige gevallen weten we dat wel De vraag of de interne representatie dat zout oplosbaar is over zout gaat, aldus Fodor,is afhankelijk van de vraag of hij over NaCl gaat. Je moet kunnen zeggen waar 'zout' naar verwijst, en dat hangt af van wat de wetenschappen daar uiteindelijk over zeggen Naturalistische psychologie zoekt naar wetten van de vorm: 'de representatie die A heeft van 'zout' verwijst naar zout dan en slechts dan als A causale relatie R heeft tot --' Omdat dit een wet moet zijn moet voor -- ingevuld worden een

karakterisering van zout die voorkomt in het vocabulaire van een wetenschap Of, in behavioristische termen, men zoekt naar wetten van de vorm '-- is de discriminatieve stimulus voor uitingen van 'zout''. -- moet een beschrijving zijn voor zout in het vocabulaire van een wetenschap, anders is bovenstaande formulering geen wet Voor zout lukt dat· NaCl is een predikaat uit de chemie. Maar verder moet de naturalistische psychologie wachten tot alle wetenschappen voltooid zijn, dat wil zeggen, tot er beschrijvingen zijn in een wetenschappelijke vocabulaire van *alles* waar we naar kunnen verwijzen.

Fodor concludeert dat in de praktijk alleen rationele psychologie mogelijk is: alleen een computationele psychologie die zich houdt aan een methodologisch solipsisme. Natuurlijk is methodologisch solipsisme geen solipsisme *tout court.* Fodor wil niet beweren dat we in dezelfde situatie verkeren als Winograd's computer. Hij weet niet wat voor relatie het is tussen hem en Robin Roberts die het mogelijk maakt om naar RR te refereren en over hem te denken Hij betwijfelt de praktische mogelijkheid van een wetenschap die over die relaties gaat. Maar hij betwijfelt niet dat die relaties er zijn, of dat hij soms echt aan RR denkt. Hij eindigt zijn artikel met de woorden:

> "My point, then, is *of course* not that solipsism is true; it's just that truth, reference and the rest of the semantic notions aren't psychological categories. What they are is: they're modes of *Dasein.* I don't know what *Dasein* is, but I'm sure that there's lots of it around, and I'm sure that you and I and Cincinnati have all got it What more do you want?" (Fodor 1980a, 71).

4.5. *Drie problemen rond Fodor's representaties.*

Fodor's representationele en computationele theorie van het mentale analyseert mentale toestanden als relaties tot interne representaties, en mentale processen als formele operaties die gedefinieerd zijn over die interne representaties Zijn theorie beschrijft bovendien hoe mentale

processen zouden kunnen verlopen - namelijk als computaties - en hoe mentale toestanden een rol kunnen spelen in de veroorzaking van gedrag - namelijk door middel van de formele eigenschappen van de interne representaties Wat zijn theorie niet beschrijft of analyseert of verklaart is de vraag wat bepaalde interne structuren in een systeem tot *representaties* maakt Fodor is zich van deze tekortkoming van zijn theorie zeer wel bewust, en probeert deze ook niet te verbergen, maar juist boven water te halen In de cognitiewetenschap wordt de notie van 'interne representatie' als vanzelfsprekend en onproblematisch gebruikt in een, naar men beweert, *fysicalistische* theorie Fodor laat juist zien dat die notie niet onproblematisch is Toch is zijn presentatie van dit punt nog erg duister, en zijn de discussies eromheen buitengewoon moeilijk en ingewikkeld

Ik denk dat de zaken op dit punt verhelderd kunnen worden wanneer we het probleem van de interne representaties opsplitsen in drie (natuurlijk nauw verwante, maar toch) verschillende problemen. Fodor heeft het over een probleem, de filosoof Robert Cummins onderscheidt in dit verband twee problemen, een semantisch probleem en een intentionaliteitsprobleem (Cummins 1983), en ik wil het semantische probleem nog eens opsplitsen en spreek van drie problemen· het referentieprobleem, het betekenisprobleem en het intentionaliteitsprobleem

4 5 1 Het referentieprobleem.

Fodor noemt zijn probleem in verband met interne representaties vaak het probleem van de relatie tussen denken (of interne representaties) en de wereld Hij is zich ervan bewust dat zijn theorie van propositionele attitudes als relaties tot interne representaties het gevaar loopt *echt* solipsistisch te zijn.

Aan het eind van zijn *The language of thought* signaleert hij dit gevaar Men zou immers uit zijn theorie het beeld kunnen krijgen van de geest die opgesloten is in een netwerk van representaties, en nooit in contact kan komen met de wereld Als ik denk aan GV (aan wie ik meer denk dan aan Fodor's RR), ben ik dan alleen gericht op mijn representatie van GV, en niet op hemzelf? Fodor doet evenwel

geruststellend over dit schrikbeeld van solipsisme:

> "To assume that mental states are analyzable as relations to representations is not to preclude the likelihood that they are *also* analyzable as relations to objects in the world. On the contrary, in the epistemologically normal situation one gets into relation with a bit of the world precisely *via* one's relation to its representation; in the normal situation, if I am thinking about Mary then it's *Mary* I am thinking about ... So there's no principled reason why a representational theory of the mind need degenerate into solipsism" (Fodor 1975, 204).

Hij voegt eraan toe dat de keten van gebeurtenissen van stimulus tot respons typisch een *causale* keten is, en dat dus ook de relatie tussen distale stimulus en proximale representatie causaal is. Als dat zo is kan zijn theorie geen solipsisme impliceren: "... there are no effects of things that aren't there" (Fodor 1975, 204).

Deze uitspraak staat evenwel op gespannen voet met de intensionaliteit van toeschrijvingen van propositionele attitudes. Daarbij is immers existentiële generalisatie niet toegestaan mensen kunnen best meningen (of wensen enz.) hebben over niet bestaande objecten. In dat geval zijn er wel interne representaties die niet het gevolg van een bestaand object zijn. Fodor lijkt die spanning evenwel niet op te merken (ik kom hier in 4.6 4 op terug)

In de uitwerking van de computationele theorie van het mentale als een theorie die onderworpen is aan de formaliteitsconditie wordt de suggestie van solipsisme nog sterker. Formele operaties hebben geen toegang tot de semantische eigenschappen van representaties. Mentale processen hoeven zich ook niets aan te trekken van de waarheid of de referentie van interne representaties. Mijn mening dat het regent maakt dat ik mijn paraplu meeneem. Dat is het geval, ongeacht of het nu waar is dat het regent of niet. De formele mentale processen die leiden tot mijn gedrag verlopen ongeacht de waarheid van mijn mening; ze hebben dan ook volgens de formaliteitsconditie geen toegang tot die eigenschap van waarheid. De interne representatie 'het regent' heeft dezelfde vorm, of hij nu waar is of niet.

In zijn solipsisme-artikel geeft Fodor goede argumenten waarom een theorie van mentale veroorzaking moet afzien van waarheid en referentie als eigenschappen van interne representaties. Zo'n theorie heeft een opake taxonomie van mentale toestanden nodig en moet dus *methodologisch solipsistisch* zijn; de theorie moet afzien van het bestaan van alles en iedereen buiten degene wiens mentale toestanden en processen en gedrag bestudeerd worden. Maar opnieuw legt Fodor er de nadruk op dat zo'n theorie van mentale veroorzaking geen *echt solipsisme* impliceert: natuurlijk kan een organisme volgens hem niet bepaalde mentale toestanden hebben als er niet een bepaalde wereld bestaat. En bovendien, als hij aan RR denkt refereert hij echt naar RR, naar iemand in de wereld. Er zijn relaties tussen hem en RR die dat refereren mogelijk maken. Fodor's probleem is dat hij een wetenschap die die relaties tot onderwerp heeft vooralsnog onmogelijk acht.

Maar in datzelfde solipsisme-artikel beveelt Fodor niet alleen een opake taxonomie van mentale toestanden en een methodologisch solipsisme aan, maar ook een formaliteitsconditie. We hebben gezien dat formaliteit wel een opake taxonomie garandeert, maar dat omgekeerd een opake taxonomie nog geen formaliteit impliceert. Door de onduidelijke argumentatie voor de formaliteitsconditie dreigt een tweede probleem rondom de representaties in het gedrang te komen: het probleem hoe representaties überhaupt betekenis kunnen hebben.

Fodor is geneigd het referentieprobleem en het betekenisprobleem als één en hetzelfde probleem te zien: het probleem van de semantische interpretatie van interne representaties. Maar wanneer hij het heeft over de praktische onmogelijkheid van een wetenschap over de relaties tussen wereld en representaties heeft hij het alleen over het referentieprobleem, en niet over het hele semantische probleem, laat staan over het intentionaliteitsprobleem.

4.4.2. *Het betekenisprobleem.*

Fodor's formaliteitsconditie garandeert dat binnen één systeem formeel identieke representaties functioneel/causaal dezelfde rol hebben. Daarmee zou je kunnen zeggen dat dezelfde vorm van een representatie

steeds samengaat met dezelfde betekenis. Maar het is nog maar de vraag of de vorm alleen kan vastleggen om *welke* betekenis het gaat De computationele theorie van het mentale heeft niet alleen niets te zeggen over de relatie tussen interne representaties en de buitenwereld, ze gaat over processen die ongevoelig zijn voor de betekenis van interne representaties Fodor laat zien dat de formaliteitsconditie een opake taxonomie van mentale toestanden garandeert, omdat formele operaties

> "... have no access to the *semantic* properties of such representations, including the property of being representations *of the environment.* (Fodor 1980a, 65).

Maar hij zegt *ook:*

> "Formal operations are the ones that are specified without reference to such semantic properties of representations as, for example, truth, reference and meaning" (Fodor 1980a, 64).

Fodor doet alsof dit hetzelfde is, maar een opake taxonomie van mentale toestanden moet wel afzien van waarheid en referentie en de omgeving, maar moet dat juist *niet* doen ten aanzien van de *betekenis* van interne representaties.

Natuurlijk is Fodor niet zo dom dat hij dat helemaal niet gezien heeft Zoals hij zegt in antwoord op eén van zijn critici: "I knew I'd get into trouble about meaning". Maar toch is hij geneigd onvoldoende onderscheid tussen betekenis en referentie te maken. Zo zegt hij dat hij *het* probleem voor representationele theorieen ziet als de vraag:

> "...what relates internal representations to the world? What is it for a system of internal representations to be semantically interpreted?" (Fodor 1980a, 203).

- alsof die twee vragen hetzelfde zijn Het lijkt er soms op alsof Fodor denkt dat alleen de referentie van een uitdrukking de semantische interpretatie bepaalt (zoals wanneer hij spreekt over semantische

interpretatie als geheel onafhankelijk van functionele rol (1980a, 106), zie ook 4.6.2). Maar dat kan niet waar zijn: uitdrukkingen met dezelfde referent in de wereld kunnen een verschillende semantische interpretatie hebben, een verschillende betekenis. En uitdrukkingen zonder een referent in de wereld kunnen een semantische interpretatie hebben, kunnen een betekenis hebben. Het is dus niet zo dat, zoals Fodor soms suggereert, een indeling van interne representaties die afziet van verschillen in waarheid en referentie en van de wereld, *ipso facto* een indeling is die afziet van de *interpretatie* van de representaties Een indeling van representaties op *formele* eigenschappen daarentegen ziet wel af van de interpretatie van die representaties.

Waarom is de semantische interpretatie, de betekenis, van interne representaties een probleem? Dat zit zo: volgens de computationele theorie van het mentale zijn mentale processen formele operaties die gedefinieerd zijn over interne representaties. Deze theorie is een fysicalistische theorie. representaties spelen een rol in de veroorzaking van het gedrag op grond van hun vorm. De representaties moeten objecten zijn met een bepaalde fysische vorm, Fodor stelt zich een soort neurale code voor Maar aan een object van een bepaalde fysische vorm valt niet af te lezen wat de semantische interpretatie is, en zelfs niet òf er een semantische interpretatie is. Hofstadter laat in zijn *Gödel, Escher, Bach* (1979) op een bladzijde tien verschillende vormen van schrift zien en merkt daarover op: "In form, there is content" (Hofstadter 1979, 169). Maar dat laat hij nu juist niet zien. De notie *dat* deze fraaie geometrische vormen inderdaad allemaal *schrift* vormen en niet alleen maar geometrische versieringen zijn, is afkomstig van totaal andere overwegingen dan van het in ogenschouw nemen van de vorm. We komen alleen op de gedachte dat zo'n reeks vormen een schrift kan zijn, met inhoud, omdat we weten dat mensen allerlei soorten schrift hebben uitgevonden, en allerlei vormen inhoud hebben toegedacht. De vorm alleen vertelt ons dat niet, en heeft die inhoud niet van zichzelf. En ook waar we vrijwel zeker weten met echt schrift van doen te hebben, zoals in het Etruskisch, waar we zelfs de vorm goed kennen - Romeins schrift - legt de vorm niet de semantische interpretatie vast De vorm heeft juist niet intrinsiek inhoud We zouden een schrift kunnen bedenken in de vorm van sneeuwkristallen,

maar dat zou niet impliceren dat er s winters boodschappen uit de hemel komen vallen De inhoud van al die schriftvormen is louter conventioneel toegekend, door mensen voor mensen gemaakt Fodor merkt in een artikel, getiteld 'Computation and cognition' op

> " *if* there is good reason for treating (some) inscriptions as linguistic tokens, then there is *equally* good reason for treating (some) neurological states as linguistic tokens" (Fodor 1981a, 174)

Maar dat lijkt me een voorbeeld van Fodoriaanse bluf, en Fodor laat zich in een interview uit 1983 ook heel anders uit op dit punt, wanneer hij zegt dat gedachten hun inhoud niet conventioneel moeten hebben maar absoluut (Miller 1983, 90) (38)

Misschien, zo zou men kunnen zeggen, kan vorm alleen niet de semantische interpretatie vastleggen, maar wordt de interpretatie wel vastgelegd als we een heel systeem van tekens hebben met alle onderlinge relaties (zie 4 6 2) Maar ook dat biedt geen uitweg Wiskundig valt te bewijzen dat ieder formeel systeem, als het een interpretatie toelaat, oneindig veel interpretaties toelaat Rey (1980) geeft, in een kritiek op Fodor's solipsisme-artikel, het volgend voorbeeld een computer wordt op verschillende dagen voor verschillende doeleinden gebruikt op woensdag houdt hij zich bezig met de SALT-onderhandelingen, op donderdag speelt hij schaak met Bobby Fischer Nu is het mogelijk dat de machine op beide dagen type-identieke computationele en fysische toestanden doorloopt Daarvoor is het nodig dat de input-decks type-identiek zijn, en dat kan als beide probleemdomeinen isomorf geconstrueerd zijn Qua vorm valt er dan niet te onderscheiden tussen de computer s mening dat Brezjnev een duik in de Donau zal nemen en zijn mening dat Fischer spoedig zal rocheren Voor het computationeel *gedrag* van de computer maakt het ook niet uit (39) Fodor erkent dit probleem (40) en zegt dat hij daarom in zijn theorie de kwestie van interpretatie (voorlopig) laat rusten. Zijn theorie van het mentale *vooronderstelt* dat de interne representaties van mensen geinterpreteerd zijn, maar verklaart niet wat het is voor een interne representatie om geinterpreteerd te zijn, of wat die interpretatie vastlegt

4.5.3. Het intentionaliteitsprobleem.

Volgens Fodor is *het* probleem voor een representationele theorie van het mentale de semantische interpretatie van de formele, fysisch bestaande, interne representaties, en de relatie tussen interne representaties en de wereld. In de vorige paragraaf heb ik laten zien dat hij het referentieprobleem en het betekenisprobleem niet goed onderscheidt. Maar er is nog een derde probleem dat Fodor ook samensmelt met zijn semantische probleem: het intentionaliteitsprobleem.

Fodor onderkent dat probleem niet als een apart probleem. Zo spreekt hij van intentionaliteit als 'semanticity' (Fodor 1980b, 431), en zegt hij in 1985:

> "... the problem of the intentionality of the mental is largely - perhaps exhaustively - the problem of the semanticity of mental representations. But of the semanticity of mental representations we have, as things now stand, no adequate account" (Fodor 1985, 99).

Die twee problemen zijn evenwel *niet* hetzelfde. Want zelfs al was het probleem opgelost wat de interpretatie van interne representaties vastlegt op unieke wijze, dan nog kan men vragen *voor wie* die representaties geïnterpreteerd zijn, *voor wie* ze iets, hoe uniek ook, representeren.

Representaties zijn immers altijd representaties *voor iemand*. Als straks de N-bommen gevallen zijn, of we allemaal aan stralingsziekte omgekomen zijn en al het leven op aarde vernietigd is, zullen onze bibliotheken en machines nog bestaan. Sommige schaakmachines die nog aanstonden zullen nog een tijdlang schaakzetten genereren, en SHRDLU bouwt nog zijn laatste toren. Maar wordt er nog geschaakt en met blokken gespeeld? Wanneer SHRDLU over zijn niet bestaande blokken 'redeneert', zijn wij het die zijn interne representaties interpreteren als representaties van blokken. Niet alleen zijn formele computaties ongevoelig voor verschillen in interpretatie van de representaties waarover ze gedefinieerd zijn, het maakt voor de computaties niets uit of die representaties überhaupt voor het systeem geïnterpreteerd zijn. Het hele proces van input, computaties en output kan plaatsvinden met

behulp van een ongeïnterpreteerd formalisme; de interne representaties hoeven *helemaal niets* te representeren. Hier ligt een parallel met een probleem dat Fodor zelf signaleerde in verband met qualia. Het functionalisme zou moeten zeggen dat iemand pijn heeft zelfs al voelt zij helemaal niets (zie 3.4.2. en 3.4.3). Maar het functionalisme samen met een computationele theorie van het mentale zou moeten zeggen dat iemand zich gedraagt en denkt en redeneert zelfs al representeren haar interne representaties voor haar *helemaal niets.*

Het functionalisme kan het verschil niet aangeven tussen een systeem met en een systeem zonder qualia; en het functionalisme samen met de computationele theorie van het mentale kan het verschil niet aangeven tussen een systeem waarvoor de interne representaties geïnterpreteerd zijn en een systeem waarvoor ze ongeïnterpreteerd, leeg zijn. De sceptische tegenwerping dat wij dat verschil bij anderen ook niet kunnen ontdekken is wat al te verificationistisch. We kennen het verschil uit ons eigen geval, en men mag van een theorie van het mentale toch verwachten dat zij niet alleen voor anderen opgaat. Zo'n theorie hoeft misschien niet direct het probleem van *other minds* op te lossen, maar het gaat te ver om te stellen dat een verschil dat *bij anderen* niet aan te tonen zou zijn, daarom ook niet bestaat (Wittgensteiniaanse argumenten over privé- taal ten spijt).

Het bekendste argument dat wij een verschil kennen tussen voor ons geïnterpreteerde en ongeïnterpreteerde representaties dat een computationele theorie van het mentale niet kan uitdrukken, stamt uit een artikel van Searle uit 1980: 'Minds, brains and programs'. Daarin geeft hij het volgende gedachtenexperiment. Stel, ik ken geen Chinees. Nu ga ik een programma à la Schank en Abelson (1977) uitvoeren. Dat gaat als volgt: Ik zit in een gesloten kamer. Ik krijg een stapel papier met Chinese karakters. Wat mij betreft kunnen het ook Japanse zijn, of fantasiekrabbels, ik zou het verschil niet merken. Een tweede stapel komt binnen. Voor mij even nietszeggend. Ik zie alleen verschillende vormen. Dan krijg ik een stapel papier met Nederlandse tekst. Die begrijp ik. Die zegt mij hoe ik elementen uit de eerste stapel, volgens vorm opgezocht, kan correleren met elementen uit de tweede stapel. Daarna krijg ik één regel karakters in mijn hok. Mijn Nederlandse tekst zegt mij precies wat ik moet doen, en na enig zoeken weet ik dat ik een regel andere karakters - mijn tekst geeft

aan van welke vorm - moet terugsturen naar buiten mijn hok. Zonder dat ik dat weet, is er buiten mijn hok iemand die de eerste stapel papier met karakters een *script* noemt, de tweede een *verhaal*, de stapel papier met Nederlandse tekst een *programma*, de korte binnenkomende regels *vragen* en de uitgaande *antwoorden*. En buiten mijn hok staat een geboren en getogen Chinese die meent een boeiende conversatie met mij te voeren. Alleen: ik ken geen woord Chinees, ik weet niet dat ik vragen beantwoord, ik weet niet eens dat ik met een taal bezig ben. Ik doe niets anders dan een computer: ik voer computationele operaties uit op formeel - naar de vorm - gespecificeerde elementen.

Searle's voorbeeld is niet onbesproken gebleven (ik bespreek hier niet de positieve argumenten uit zijn artikel, enkel de Chinese kamer). De meest voor de hand liggende tegenwerping is dat Searle dan wel geen Chinees begrijpt, maar de kamer-met-Searle-erin wel. Deze tegenwerping houdt geen stand want Searle kan de programma-instructies uit zijn hoofd leren en nog geen Chinees begrijpen. En de tegenwerping dat het onwaarschijnlijk is dat Searle, rondlopend-in-de-wereld met het programma in zijn hoofd, niet zou begrijpen waar zijn Chinese conversaties over gaan doet niet ter zake: Searle zou het dan vast wel snel leren, maar zijn argument is dat het instantiëren van een programma niet *ipso facto* betekent dat hij Chinees begrijpt.

Cummins (1983) oppert de mogelijkheid dat het kunnen beantwoorden van Chinese vragen een te geïsoleerde capaciteit is, dat het instantiëren van het programma los van alle andere capaciteiten daarom geen begrijpen met zich mee brengt. Misschien is er wel sprake van begrijpen als het programma "... is installed in a sufficiently ritzy neighborhood of other computational capacities, all properly integrated .. " (Cummins 1983, 108) Ik denk evenwel dat ook deze tegenwerping voorbijgaat aan waar het om gaat. Merk op dat Cummins al toegeeft dat het instantiëren van het programma op zichzelf niet betekent dat er sprake is van begrip. Cummins lijkt echter niet te zien waarom er geen begrip is. Dat begrip is er niet omdat voor Searle de Chinese karakters ongeïnterpreteerd zijn Wanneer hij zijn capaciteit voor het beantwoorden van Chinese vragen integreert met zijn andere capaciteiten zal hij snel Chinees leren, omdat hij de karakters dan kan *vertalen* in zijn eigen taal. Maar wanneer de interne representaties *in*

het Mentalees ongeïnterpreteerd zijn, is er geen voor het systeem geïnterpreteerde taal meer over om in te vertalen. Een systeem dat *alleen* maar het programma instantieert zal op de vraag "Begrijp je Chinees" ook al "Ja" antwoorden. Toevoeging van andere programma's voor andere capaciteiten zal daar niets meer aan veranderen of verbeteren. Wanneer de interne representaties voor het systeem ongeïnterpreteerd zijn, kan het louter toevoegen van meer operaties, gedefinieerd over die representaties, daar niets aan veranderen

Ook Cummins' volgende oplossing kan geen uitkomst bieden om dezelfde reden Hij stelt dat een systeem intentionaliteit heeft en echt begrijpt wat het doet als het ook de semantiek van zijn andere representaties representeert. Dan zou het systeem de betekenis van zijn *andere* representaties begrijpen, maar alleen als die semantiek *zelf* gerepresenteerd is in een *al voor het systeem geïnterpreteerde taal.* Het probleem herhaalt zich gewoon

Let wel, in Searle's voorbeeld is het feit dat Searle de niet-Chinese programma-instructies begrijpt irrelevant en verwarrend Een machine is zo gebouwd dat hij de instructies in machinetaal *uitvoert;* hij hoeft ze niet te begrijpen (zie ook 4 6 1).

Fodor is het op dit punt met Searle éens (dit, gezien de reacties op Searle's artikel, in tegenstelling tot vele andere AI-wetenschappers): het uitvoeren van formele operaties op zichzelf, het realiseren van een computerprogramma, betekent nog niet dat er sprake is van begrip, van 'weten waar het over gaat'. Of, zoals Fodor het in zijn eigen solipsisme-artikel zegt:

> If the *programmer* chooses to interpret the machine inscription "Robin Roberts won 28" as a statement about Robin Roberts (e.g., as the statement that he won 28), that's all well and good, but it's no business of the machine's. The machine has no access to that interpretation, and its computations are in no way affected by it. The machine doesn't know what it's talking about, it doesn't care; *about* is a semantic relation" (Fodor 1980a, 65)

Alleen denkt Fodor dat het intentionaliteitsprobleem kan worden opgelost als de semantische interpretatie van interne representaties kan

worden vastgelegd Hij lijkt niet goed in te zien dat zelfs als er een interpretatie *is*, er nog altijd iemand moet zijn die de interpretatie *uitvoert*, *voor wie* het een interpretatie is

4.6. Pogingen om het referentieprobleem en het betekenisprobleem op te lossen.

Lange tijd leek Fodor er vrede mee te hebben dat zijn theorie van het mentale - en de cognitieve psychologie - geen verklaring kunnen geven voor de betekenis en de referentie van interne representaties. Die betekenis wordt altijd al vooronderstelt En daarmee wordt intentionaliteit, het vermogen van personen om ergens op gericht te zijn, iets buiten zichzelf te kunnen representeren voor zichzelf, vooronderstelt.

Misschien is dat voor de psychologie niet erg Misschien gaat de cognitieve psychologie helemaal niet over wat mentale toestanden zijn, maar alleen over hoe de formele computaties gaan Cognitieve psychologie is dan een soort uitgebreide formele redeneerkunde of logica (maar zie 2 4 voor Fodor's twijfels over de formaliseerbaarheid van alle denkprocessen). In een artikel uit 1978, 'Tom Swift and his procedural grandmother', zegt Fodor te kunnen leven met het oninteressante feit dat 'stoel' verwijst naar stoelen, en dat hij geen theorie van referentie heeft en geen mechanisme om die theorie te realiseren.

Maar Fodor heeft toch graag wèl een mechanisme Hij heeft wel een mechanisme gevonden voor de relatie tussen organismen en proposities, namelijk mediering door interne representaties, maar nu heeft hij een mechanisme nodig voor de relatie tussen interne representaties en proposities, een theorie die verklaart wat het is voor een representatie om semantisch geinterpreteerd te zijn en een mechanisme om die theorie te kunnen realiseren.

Fodor weet dat zijn theorie van het mentale nog steeds intentionaliteit vooronderstelt. Maar waar hij uiteindelijk op uit is is een fysicalistische theorie van het mentale, en een fysicalistische oplossing voor het lichaam-geest probleem. Dus de gezochte theorie die moet verklaren wat het is voor een representatie om semantisch

geïnterpreteerd te zijn, en wat die interpretatie vastlegt, moet uiteindelijk ook fysicalistisch zijn. Zo'n theorie mag niet in de explanans weer verwijzen naar een interpretator of begrijper, of naar intentionaliteit. Fodor is op zoek naar een theorie van de algemene vorm *'R representeert S' is waar dan en slechts dan als C*, waarbij *R* staat voor een representatie in de zin van een formeel, fysisch object, *S* voor datgene wat *R* representeert, en *C* voor een specificatie van eigenschappen van en relaties tussen *R* en *S* die samen voldoende en noodzakelijke voorwaarden vormen voor het feit dat *R S* representeert. *C* mag daarbij zelf niet verwijzen naar representaties, of naar intentionaliteit 'Mijn representatie van GV representeert GV omdat ik er GV mee bedoel' is dus niet het soort verklaring waar Fodor naar op zoek is. De voldoende en noodzakelijke voorwaarden *C* moeten gesteld worden in een fysicalistische taal.

Zo'n theorie van de semantische interpretatie van interne representaties zou heel fundamenteel zijn. Volgens Fodor zijn de semantische eigenschappen van uitdrukkingen in de natuurlijke taal afgeleid van die van de propositionele attitudes, en de semantische eigenschappen van propositionele attitudes zijn afgeleid van die van de interne representaties Dus een theorie die verklaart waar de interne representaties hun semantische eigenschappen vandaan hebben mag niet meer verwijzen naar taal of propositionele attitudes, of naar intentionaliteit.

Fodor zegt dat zo'n theorie er niet is. Hij bespreekt wel voorstellen voor zo'n theorie die hij afwijst. In 1980 wil hij laten zien waarom zo'n theorie praktisch onmogelijk is Maar daarna probeert hij toch voorzichtig een paar keer zo'n theorie te formuleren, al zegt hij in 1985 nog steeds· "At best, however, it's a long way off" (Fodor 1985, 99).

In 4.6.1 en 4.6.2 geef ik Fodor's argumenten om respectievelijk de procedurele semantiek en de functionele-rol-semantiek af te wijzen als een voorgestelde oplossing voor het probleem van de semantische interpretatie van interne representaties Ik stem grotendeels met zijn argumenten in en zal proberen ze soms nog te versterken In 4.6.3 en 4.6.4 bespreek ik Fodor's eigen pogingen tot een theorie van semantische interpretatie, en wil ik laten zien dat ook zijn oplossingen niet houdbaar zijn.

4.6.1. Procedurele semantiek.

Het basisidee van de procedurele semantiek is het volgende: een computer bevat zeker geen geest in de machine, en werkt met interne representaties in een interne taal die niet verder geïnterpreteerd hoeft te worden. Er is noch sprake van oneindige regressie, noch van dualisme. In een computer wordt de programmeertaal vertaald (*interpreted* of *compiled*), eventueel via een aantal tussenstappen, in de machinetaal De machine is zo gebouwd dat hij de instructies in de machinetaal zonder meer uitvoert, dat wil zeggen, zonder dat die instructies zelf weer gelezen en geïnterpreteerd hoeven te worden. Dit idee staat centraal in het beroemde werk van Winograd over SHRDLU. Zijn programma opereert op een input-zin om een representatie van de betekenis van die zin te produceren in een interne taal (Winograd 1971). Die interne taal heeft de vorm van instructies voor procedures; zijn boek heeft ook als titel *Procedures as a representation for data in a computer program for understanding natural language*. En Johnson-Laird zegt over de procedurele semantiek het volgende:

> "...artificial languages which are used to communicate programs of instructions to computers, have both a syntax and a semantics. Their syntax consists of rules for writing well-formed programs that a computer can interpret and execute. The semantics consists of the procedures that the computer is instructed to execute" (Johnson-Laird 1977, 189);

en

> "...we might speak of the intension of a program as the procedure that is executed when the program is run" (Johnson-Laird 1977, 192).

Ofschoon Fodor in *The language of thought* uit 1975 het begrijpen van zinnen uit een natuurlijke taal ziet als het vertalen ervan in een ambiguïteits-vrije interne taal, het Mentalees, verwerpt hij daar al, in een voetnoot, de procedurele semantiek. In 1978 wijdt hij er een heel

artikel aan, getiteld 'Tom Swift and his procedural grandmother (herdrukt in Fodor 1981a) Zo mogelijk nog opgewekter dan anders gaat hij in dit artikel de procedurele semantiek te lijf met argumenten die er niet om liegen

In zekere zin is het waar dat programmeertalen (die steeds meer gaan lijken op natuurlijke talen) vertaald worden in de machinetaal, en dat de machinetaal zelf een geinterpreteerde taal is Maar in een andere zin geeft de machinetaal, via de compilatie, geen interpretatie van de programmeertaal De interpretatie die de machinetaal geeft aan een zin uit de programmeertaal is normaliter niet de bedoelde interpretatie, het is normaliter niet de interpretatie die specificeert *wat de zin betekent*. Machines weten niet waar de programma's die ze draaien over gaan, alles wat ze 'weten is hoe ze de programma s moeten draaien (Hier gebruikt Fodor het bovengenoemde voorbeeld van Georges Rey over een computerprogramma dat geinterpreteerd kan worden op twee manieren als SALT-onderhandelaar en als schaker) Fodor bespreekt een voorbeeld dat gegeven wordt door Miller en Johnson-Laird (1976) van hoe een procedureel systeem omgaat met de vraag 'Heeft Lucy het toetje gebracht?' De vraag wordt vertaald in instructies volgens welke het geheugen wordt afgezocht naar herinneringen aan Lucy Deze herinneringen worden dan nageplozen op verwijzingen naar (bijvoorbeeld) chocoladecake, en hoe het systeem de vraag beantwoordt wordt bepaald door dit soort verwijzingen

Maar, zegt Fodor, de auteurs hebben het recht niet te spreken over herinneringen aan Lucy en verwijzingen naar chocoladecake Want niets in hun theorie kan de relaties reconstrueren tussen Lucy' en Lucy of tussen chocoladecake en chocoladecakes Alles wat in de machine-taal voorkomt is een instructie om te gaan naar het *adres* met als label 'Lucy' (maar dat net zo goed - en minder misleidend - als label '#959 kon hebben (mogelijk ook heeft)) en te zien of daar een formule voorkomt die pas in de compilatie gepaard gaat met het predikaat 'brengt een chocoladecake'

Machinetaal is geinterpreteerde taal, maar de interpretatie geeft als denotatum voor 'Lucy' niet Lucy-het-meisje maar enkel Lucy-het-adres De interpretatie van de machinevertaling gaat niet over de vraag of Lucy het toetje bracht, maar over de vraag of een bepaalde formule op een bepaald adres voorkomt

Fodor noemt een analoog geval. de vraag 'heeft Napoleon de slag bij Waterloo gewonnen?' wordt semantisch geinterpreteerd als. *zoek uit of de zin "Napoleon heeft de slag bij Waterloo gewonnen" voorkomt in het boekdeel met Dewey decimaal nummer XXX, XXX in het 42e straat filiaal van de New York City Public Library.* En hij voegt daar, met onbedwingbaar vermaak, aan toe.

> "'But', giggled Granny, 'if that was what 'Did Napoleon win at Waterloo' meant, it wouldn't even be a question about *Napoleon'*. 'Aw, shucks' replied Tom Swift" (Fodor 1981a, 210-211).

Ik denk dat Fodor de procedurele semantiek terecht afwijst, maar zijn argument is hier wel iets te gemakkelijk, het *begs the Frege* zogezegd (zie Dennett 1973) Dat komt omdat hij het onderscheid tussen betekenis en referentie (of tussen intensie en extensie, als we de beide begrippen als verwisselbaar beschouwen) niet goed maakt. Nogmaals, volgens Fodor is semantische interpretatie van interne representaties een kwestie van referentie, en een kwestie van de (causale) relatie tussen interne representaties en hun referenten in de buitenwereld En natuurlijk, er zijn geen directe relaties tussen het *adres* met als label 'Lucy' in de machine en het *meisje* Lucy Maar Johnson-Laird heeft ook nergens beweerd dat die relaties er zijn. Wat hij zegt is dat de programmaprocedures de *intensie,* de betekeins, van de programmatermen vastlegt Om die claim te weerleggen zou Fodor een ander argument moeten geven.

Het probleem kan nog in andere bewoordingen worden uitgedrukt. Fodor laat duidelijk zien dat de machine niet *over* Lucy kan menen dat ze het toetje heeft meegebracht; de machine kan niet een zogenaamde *de re* mening hebben, om de eenvoudige reden dat de machine de *res* in kwestie, Lucy, niet kent - de machine kent geen enkele *res.* Maar misschien kan de machine wel *de dicto* menen dat Lucy het toetje heeft meegebracht. Het verschil tussen *de re* en *de dicto* meningen wordt meestal uitgedrukt in de manier van toeschrijven. Als ik meen dat de burgemeester van Tsjernobyl een partijmarionet is dan is dat een *de dicto* mening; er is niemand, geen concreet persoon, geen object in de wereld over wie ik dat meen, want ik weet helemaal niets van de

burgemeester van Tsjernobyl, als die al bestaat Maar als ik vind dat de partijleider van Rusland meer informatie zou moeten geven, dan is dat wel een *de re* mening, een mening *over* Gorbatsjov; ik meen dan over Gorbatsjov dat hij meer informatie moet geven.

In Fodor's theorie wordt dit verschil tussen *de re* en *de dicto* meningen niet gemaakt. Propositionele attitudes zijn in zijn theorie altijd analyseerbaar als relaties tot interne representaties, en in die zin dus altijd *de dicto,* over een zin. Maar ze zijn, normaliter, ook altijd analyseerbaar als relaties tot objecten in de wereld (Fodor 1975, 204), en in dat opzicht, normaliter, altijd *de re,* over een object (zie 4 6.4 voor problemen rond de notie 'normaliter'). Merk op dat het onderscheid *de re - de dicto* niet hetzelfde is als het onderscheid transparant - opaak. Het onderscheid transparant - opaak bestaat alleen vanuit het standpunt van de buitenstaander, de beschrijver van andermans mentale toestanden (zie 4.3). Oedipus' mening dat hij met Iokaste is getrouwd is bij een opake indeling verschillend van zijn mening dat hij met zijn moeder is getrouwd, bij een transparante indeling dezelfde. Maar het zijn allebei *de re* meningen, meningen *over* de persoon Iokaste, zijn moeder. Oedipus' mening dat de veroorzaker van de pest gestraft moet worden is aanvankelijk *de dicto,* zolang hij niet weet wie die veroorzaker is, en pas op het eind van het verhaal *de re,* over hemzelf (voor meer gedetailleerde besprekingen over het *de re - de dicto* onderscheid zie Woodfield 1982).

Fodor geeft geen expliciet argument dat de door Johnson-Laird beschreven machine geen *de dicto* meningen kan hebben, en dat de procedurele semantiek niet de betekenis (de intensie) van de programmeertaal vastlegt. Hij stelt dat de interpretatie van de machinevertaling niet de bedoelde interpretatie is, en ik ben het daarmee eens. Maar de reden kan niet zijn dat er voor de machine geen relatie bestaat tussen het adres met als label 'Lucy' of 'chocoladecake' en Lucy-het-meisje en chocoladecake. Zo'n relatie is er voor mij evenmin. Mijn meningen over Lucy en haar chocoladecake zijn al evenmin *de re,* maar mijn interne representaties hebben wel een betekenis. De reden waarom de machinevertaling niet de bedoelde interpretatie geeft is dat de machine überhaupt nooit in contact is geweest met de buitenwereld. Voor het vastleggen van de interpretatie van interne representaties is het niet nodig dat er voor iedere

representatie een relatie tot een referent in de buitenwereld is, maar enkel dat er een veel globalere relatie tot de buitenwereld bestaat (of tenminste, bestaan heeft).

Fodor geeft aan het einde van zijn solipsisme-artikel een hint in deze richting wanneer hij stelt dat het vastleggen van de semantische interpretatie van interne representaties een kwestie is van *Dasein*, van in-de-wereld-zijn (41). Hij gaat er echter nergens op door, en spreekt elders steeds van de specifieke relaties tussen representaties en de wereld. Hij wil met zijn methodologisch solipsisme afzien van die specifieke relaties tot referenten in de buitenwereld, maar blijft onduidelijk over de existentiele *commitments* die hij wil maken. Zo geeft hij toe dat:

> " .. such ascriptions as "it seems to S as though there were food" in *some* sense presuppose the existence of food; if not there then, at least somewhere sometime" (Fodor 1980a, 101).

Maar dat is volgens hem omdat de *toeschrijving* zelf gebeurt in een taal die alleen gesproken kan worden als er een wereld bestaat, zodat er voor de *toeschrijver* ooit ergens voedsel moet bestaan. Maar zelfs als de toeschrijving opaak is, zoals dat volgens Fodor moet, dan moet ook voor de *toegeschrevene* ooit ergens voedsel bestaan hebben. Die persoon zou nooit zelfs maar kunnen hallucineren dat er voedsel was, als ze niet ooit in contact met voedsel geweest was.

Dit punt is door sommige filosofen verdedigd als de positie dat tenminste een aantal gedachten en meningen altijd *de re* zijn. Die gedachten en meningen zouden niet kunnen bestaan als de bijpassende objecten in de wereld niet bestonden Wat *in* het organisme gebeurt kan niet helemaal de betekenis van de interne representaties vastleggen Het bekendste argument voor zo'n positie is afkomstig van Putnam (1975, zie ook Burge 1979, 1982, Woodfield 1982). Het is een gedachtenexperiment waarbij we ons een Tweelingaarde moeten voorstellen ergens in het heelal, die slechts in één opzicht van onze aarde verschilt wat onze dubbelgangers daar 'water' noemen is geen H_2O maar een geheel andere chemische substantie (Het is nogal ongeloofwaardig dat zo'n verschil niets zou uitmaken - we *bestaan* voor het grootste deel uit water - maar laten we niet moeilijk doen.)

Kunnen de mensen daar, met dezelfde gesproken taal als wij en formeel dezelfde interne representaties, een mening over *water* hebben? De algemene intuïtie is dat ze dat niet kunnen, omdat noch zij, noch hun voorouders ooit in contact met water zijn geweest. Net zomin als een lid van een Papoeastam die nog nooit contact met een andere cultuur heeft gehad een mening kan hebben dat het postkantoor om half zes sluit, ook al heeft hij in zijn hoofd formeel dezelfde structuren als ik nu.

Welke propositionele attitudes iemand kan hebben, is niet alleen afhankelijk van wat er in zijn hoofd gebeurt, maar ook van de aard van zijn fysische en sociale omgeving. Dat is ook waar als we zijn propositionele attitudes opaak toeschrijven. Wanneer Macbeth denkt een dolk te zien hoeft er niet nu en hier een dolk te zijn. Maar het is niet mogelijk dat Macbeth dat denkt wanneer hij noch direct, noch via anderen (leraren, boeken) ooit met een dolk in contact was geweest. De aard van de omgeving bepaalt mede welke betekenis de interne representaties kunnen hebben. En als er, in het geval van computers, geen omgeving is, of geen enkel contact met een omgeving, dan hebben de interne representaties geen andere betekenis dan wat door de interne machine-omgeving wordt vastgelegd. Dan geeft de machinetaal niet de bedoelde interpretatie van de termen van de programmeertaal. Het is de afwezigheid van relaties met een omgeving überhaupt, en niet, zoals Fodor denkt, de afwezigheid van meer directe referentierelaties, die maakt dat de machinetaal niet de juiste semantische interpretatie vastlegt.

Een tweede argument van Fodor tegen de procedurele semantiek, namelijk dat deze verificationistisch is, treft wat meer zijn doel. Een probleem van procedurele semantiek is dat zinnen geïnterpreteerd worden door middel van procedures. Nu is het misschien nog niet zo merkwaardig om een vraag te interpreteren als een procedure (om een antwoord te vinden). Maar bij zinnen in de stellende wijs moet de interpretatie ook een instructie of een procedure zijn; immers, alle zinnen worden in de machinetaal vertaald in procedures om iets te doen. Die procedure zou kunnen zijn om de aangeboden informatie in het geheugen op te slaan. Maar meestal wordt daaraan vastgeknoopt een controle of de informatie niet strijdig is met eerder ontvangen informatie, en volgens sommigen vormt deze laatste procedure de

eigenlijke interpretatie van een zin in de indicatief

Zo zegt Woods dat, wil een intelligent wezen de betekenis van atomaire zinnen weten, het een verzameling effectieve criteria moet hebben om uit te maken of de zin waar of onwaar is (Woods 1975, 39) Dat is echter wel heel erg verificationistisch volgnes Fodor Het klassieke verificatie-criterium voor betekenis stelde enkel dat er voor een betekenisvolle zin de logische mogelijkheid voor een verificatiemethode bestond Hij merkt op

> "Good grief, Tom Swift, if all of us English speakers know how to tell whether positrons are made out of quarks, why doesn't somebody get a grant and find out?" (Fodor 1981a, 216)

Ook hier is Fodor evenwel wat weinig genereus voor zijn tegenstanders Woods' uitspraak is zo ongelukkig dat men haast mag aannemen dat hij het zo niet bedoeld kan hebben Wat de procedurele semantiek beweert is dat de interpretatie van een zin gegeven wordt door een procedure die checkt - met een effectieve set van criteria - of die zin past in de al aanwezige kennisrepresentatie (en natuurlijk niet in iedere mogelijke wereld) Wat de problemen voor zo n effectieve procedure zijn is al besproken in 2 4 als het *frame*-probleem

Fodor heeft wel gelijk wanneer hij zegt dat zo'n procedure wellicht is wat er gebeurt als een mens een zin begrijpt - weet wat de betekenis is - maar dat daarmee nog geen semantiek is gegeven. Immers, de notie van 'weten wat een atomaire zin betekent' is dan wel geanalyseerd, maar er wordt een geheel geinterpreteerde kennisrepresentatie in die analyse vooronderstelt (zie ook Putnam 1983b) Hoe komt die aan zijn semantische interpretatie? Die vraag is nog totaal onbeantwoord

4 6 2 Functionele-rol-semantiek

In Fodor's computationele theorie van het mentale vervullen interne representaties met dezelfde vorm dezelfde functionele rol Daarmee is een opake taxonomie van mentale toestanden gegeven, ik ben in

dezelfde mentale toestand als GV wanneer we beide denken: "Ik heb hoofdpijn", ook al is de referent van 'ik' in beide gevallen niet dezelfde (nl. GV en MM). Beide interne representaties hebben dezelfde vorm en spelen dezelfde functionele rol; ze leiden ertoe dat ik aspirine in mijn mond stop en GV aspirine in de zijne.

Toch denkt Fodor niet dat de semantische interpretatie van interne representaties vastgelegd kan worden door hun functionele rol. Hij zegt (of liever, hij zegt te hopen) dat de inhoud van interne representaties bepaald wordt door hun functionele rol *en* door semantische interpretatie. Waarom vindt hij semantische interpretatie iets anders dan functionele rol? Waarom wijst Fodor een functionele-rol-semantiek (zie b.v. Loar 1981, McGinn 1982b) af?

Een procedurele, verificationistische semantiek moet in feite samengaan met een functionele-rol-semantiek, ook al wordt dat niet expliciet genoemd. De machinetaal speelt immers een functionele rol in de toestandsveranderingen van de machine; de machinetaal veroorzaakt die veranderingen. Fodor noemt evenwel de functionele-rol-semantiek niet in zijn bespreking van de procedurele semantiek en evenmin in zijn solipsisme-artikel, waarin hij functionele rol gescheiden ziet van semantische interpretatie. Zijn argumenten tegen de functionele rol semantiek zijn alleen - en zeer summier - te vinden in Fodor 1985. Ik zal een aantal argumenten tegen de functionele-rol-semantiek formuleren en aangeven op welke punten ik denk dat Fodor de argumentatie deelt.

In zijn Tom Swift-artikel laat Fodor zien dat de machinetaal niet de bedoelde interpretatie geeft, omdat de machine nooit contact gehad heeft met de buitenwereld. Hij haalt dan ook het voorbeeld aan van het computerprogramma dat op twee manieren geïnterpreteerd kan worden: als schaker en als raket-onderhandelaar. De functionele rol van de interne representaties *onderdetermineert* de semantische interpretatie. Men zou hier tegenin kunnen brengen dat die onderdeterminatie veroorzaakt wordt door de beperktheid en de geïsoleerdheid van de programma's (zie ook Cummins 1983). Een systeem dat veel meer kon, en op geïntegreerde wijze, zou niet zo multi-interpretabel zijn. Maar deze tegenwerping gaat niet op. Van een computerprogramma dat zowel kan schaken als onderhandelen op dezelfde dag valt nog steeds niet uit te maken *wanneer* het schaakt of onderhandelt. De probleemdomeinen

blijven isomorf, en voor het computationeel gedrag van de computer maakt het nog steeds niet uit of Fischer zal rocheren of Brezjnev in de Donau zal springen.

Maar als het systeem nu eens in causale relatie met de wereld staat, in die zin dat het veranderingen in de wereld veroorzaakt. De interne representaties veroorzaken dan niet meer alleen de *machine* veranderingen. Dan gaat Fodor's argument uit het Tom Swift-artikel niet meer op, want dan gaat de interpretatie van de machinetaal niet meer alleen over machineveranderingen. Maar deze oplossing brengt wel andere problemen met zich mee.

Er is het probleem van de individualisatie en de identificatie van mentale toestanden. Als de inhoud van een mentale representatie wordt vastgelegd door de functionele rol die die representatie speelt in het hele netwerk van interne representaties en hun gevolgen voor het gedrag, dan hebben twee representaties alleen dezelfde interpretatie als ze dezelfde rol spelen in hetzelfde netwerk. Net als bij het Turingmachine-functionalisme (zie 3.4.1) geeft dat een te fijne indeling van mentale representaties (zie ook Davidson 1970). Bovendien zijn er geen duidelijke criteria voor de individualisering van functionele of causale rol. Hamlet's mening dat zijn oom achter het scherm stond veroorzaakte (mede) dat hij Polonius vermoordde; Oedipus' mening dat Iokaste aantrekkelijk was veroorzaakte (mede) dat hij met zijn moeder trouwde; en Princip's voornemen om Franz Ferdinand te vermoorden veroorzaakte (mede) dat hij de eerste wereldoorlog deed ontbranden. Hun mentale toestanden speelden een functionele rol, ze veroorzaakten (mede) die catastrofen. Maar de semantische interpretatie van hun interne representaties heeft niets te maken met die rampen. Welke gebeurtenis in de eindeloze causale keten legt de semantische interpretatie van de interne representaties vast? Wanneer je de functionele rol de interpretatie laat bepalen en je blijft binnen een solipsistisch systeem, dan gaat de interpretatie alleen over machineveranderingen, dan heb je te weinig interpretatie. Maar wanneer je de veranderingen in de wereld bij de functionele rol betrekt, dan heb je te veel interpretaties: er is dan een *embarras du choix* van mogelijke interpretaties. In Fodor's theorie moet een bepaalde interne representatie echter een *bepaalde* semantische interpretatie hebben.

Je zou kunnen zeggen dat de interpretatie van een mentale toestand niet in isolatie bepaald kan worden, maar dat het netwerk van interne representaties in hun functionele interrelaties en relaties met veranderingen in de wereld in zijn geheel, wel een interpretatie vastlegt Maar, afgezien van de te fijne indeling die zo'n holisme met zich meebrengt, zou dat alleen gelden als het systeem in kwestie ideaal rationeel zou zijn De mentale toestand die een functionele rol speelt in het causale (mini)netwerk van de mening dat het regent, de wens om niet nat te worden en het meenemen van een paraplu bij het naar buiten gaan, is de mening die geinterpreteerd kan worden als "Een paraplu zorgt dat je niet nat wordt' of iets dergelijks Maar dat geldt alleen voor een rationeel persoon Voor een - naar westerse maatstaven - minder rationeel persoon zou de mening dat een paraplu de weergoden gunstig stemt dezelfde causale rol kunnen spelen Hoe minder rationeel iemand is, des te minder kan de interpretatie van zijn interne representaties vastgelegd worden door hun functionele rol (zie ook Stich 1982a, 1982b, 1983) Nu is niemand van ons volledig rationeel we maken soms verkeerde gevolgtrekkingen, we zien niet goed de consequenties van onze meningen enz Maar het is erg moeilijk een niet-arbitraire norm vast te stellen voor rationaliteit volgen de causale relaties tussen de interne representaties een bepaalde logica, en zo ja, welke? En hoeveel 'fouten' tegen die logica zijn dan toegestaan, en wanneer moeten we concluderen dat we een andere logica moeten vooronderstellen (Dennett 1978b, Fodor 1985, zie ook 4 6 4)?

Tenslotte kan men tegen een functionele-rol-semantiek dezelfde bezwaren inbrengen die in de vorige paragraaf genoemd zijn De interne representatie van onze dubbelganger op Tweelingaarde die dezelfde functionele rol speelt als onze gedachte over water kan niet een gedachte over water zijn, omdat onze dubbelganger nooit met water in contact geweest is, maar enkel met een andere vloeistof

Al deze argumenten geven aan dat functionele rol wel veel te maken heeft met semantische interpretatie, maar dat functionele rol, als puntje bij paaltje komt, de semantische interpretatie niet kan vastleggen Fodor is die mening ook toegedaan - hij noemt de problemen rond idealisatie van rationaliteit en van de individualisatie van interne representaties in zijn artikel 'Fodor's guide to mental

representations' (1985) Zijn enige uitweg om de semantische interpretatie van interne representaties vast te leggen is om de causale relaties tussen de wereld en het systeem, in die zin dat de wereld veranderingen in de interne representaties veroorzaakt, bij de interpretatie te betrekken. Merk op dat het hier om andere causale relaties gaat dan de relaties waarbij de interne representaties veranderingen in de wereld veroorzaken, die boven al bij de functionele-rol-semantiek betrokken waren.

Deze causale relaties van wereld naar systeem moeten de problemen van multi-interpreteerbaarheid en de Tweelingaarde-problemen kunnen oplossen, in Fodor's opvatting Hij meent dat naast de interne, functionele-rol-relaties tussen de interne representaties, ook en vooral (soms beweert hij zelfs alleen) de wereld een rol speelt in het vastleggen van de interpretatie van interne representaties (zie voor een twee factoren of twee componenten theorie ook b.v McGinn 1982b, Loar 1980, 1981)

In de volgende paragrafen zal ik twee pogingen van Fodor bespreken om met behulp van die relaties van wereld naar systeem het probleem semantische interpretatie op te lossen.

4.6.3 'Narrow content' en fenomenalisme.

In de bespreking van het voorbeeld van SHRDLU en van de schaker/onderhandelaar hebben we gezien dat de interpretatie van de machinetaal niet de bedoelde interpretatie van de programmataal vastlegt Computers hebben geen toegang tot de bedoelde interpretatie van het programma dat ze draaien het maakt voor de computer niet uit of zijn programma geïnterpreteerd is als een SALT-onderhandeling of als een schaakpartij. De betekenis en referentie van de functionele toestanden van een computer kunnen niet vastgelegd worden. Een computer heeft dan ook helemaal geen interactie met de buitenwereld, het is een werkelijk solipsistisch systeem Zelfs Winograd's SHRDLU heeft nog nooit een blok gezien, om de simpele reden dat het programma gedraaid wordt op een machine die geen ogen (of televisiecamera's) heeft.

Een standaard-antwoord van cognitiewetenschappers op het argument

dat de interne representaties van computers ongeinterpreteerd zijn, bestaat er dan ook uit een voorstel te doen om robots te maken computers met sensorische transducers die vrijelijk kunnen interacteren met een natuurlijke omgeving De sensorische transducers zijn gevoelig voor fysische stimuli uit de buitenwereld, net zoals onze ogen en oren Als we zo n robot hebben, dan kan niemand meer volhouden dat zijn interne representaties niet geinterpreteerd zijn Dan hebben we een mens nagemaakt, en weten we dus hoe een mens werkt Pylyshyn zegt in dit verband heel expliciet dat hij computatie ziet als een letterlijk model van mentale activiteit en niet als een simulatie van gedrag (Pylyshyn 1984, 43-44) Hij schrijft dat, als we zo n robot met transducers hebben, die vrijelijk interacteert met zijn omgeving,

> " it is far from obvious what if any latitude the theorist (who knows how the transducers operate and therefore what they respond to) would still have in assigning a coherent interpretation to the functional states " (1984, 44) (42)

Laten we eens precies bekijken hoe dit voorstel het probleem van de semantische interpretatie van interne representaties moet oplossen We gingen ervan uit dat in de machinetaal de termen (zoals Lucy' en 'toetje') niet geinterpreteerd waren, maar de opdrachten voor de machine (zoals vergelijken , 'vervangen', eventueel 'logisch afleiden') wel We zouden kunnen zeggen dat alle formele logische concepten geinterpreteerd zijn en de niet-logische concepten niet We hebben in de machinetaal te doen met een ongeinterpreteerd formeel systeem, waarvan nog vele interpretaties mogelijk zijn Nu is het voorstel om die interpretatie vast te leggen met behulp van de input van de *transducers*. De machinetaal spreekt dan niet meer alleen over machineprocessen en lege labels, maar ook over transducertoestanden Hebben we dan een interne taal die volledig en eenduidig geinterpreteerd is en bovendien de bedoelde interpretatie van het programma (ongeveer van natuurlijke taal) (43) geeft? Dat is nog maar zeer de vraag

De kwestie is als volgt voor Fodor de semantische eigenschappen van natuurlijke taal zijn afkomstig van de semantische eigenschappen van propositionele attitudes, en de semantische eigenschappen van

propositionele attitudes zijn afkomstig van de semantische eigenschappen van de interne representaties in de interne taal (het Mentalees). Als het Mentalees een soort machinetaal is, dan moet dus gelden dat de semantische eigenschappen van de interne taal volledig vastgelegd kunnen worden in termen van logische concepten en transducertermen. En dat moet inhouden dat alle termen definieerbaar zijn in termen van logische concepten en transducertoestanden.

Het probleem zit hem in de transducers, en het is een tweeledig probleem. De sensorische transducers hebben als input fysische stimuli uit de buitenwereld. En ze hebben als output een verzameling van transducertoestanden, die groter of kleiner kan zijn naar gelang de gevoeligheid van de transducer. Die transducertoestanden zijn al symbolisch, en kunnen interacteren met de andere interne symbolen van het systeem.

Men zou kunnen zeggen dat die transducer outputs sensatie- of gewaarwordingstoestanden zijn, bijvoorbeeld 'rood', of 'licht balk met die-en-die oriëntatie' (ik wil alle verschillende mogelijkheden en ingewikkeldheden uit de zintuigfysiologie hier buiten beschouwing laten, maar ik heb hier inderdaad even gedacht aan de experimenten van Hubel en Wiesel). Het probleem is dan dat, willen alle interne representaties geïnterpreteerd zijn, alle niet-logische concepten uitgedrukt moeten kunnen worden in sensatie-concepten. Hier ontmoeten we een oude bekende: het fenomenalisme, het programma om alle niet-logische concepten te reduceren tot sensatie-concepten via coordinerende definities.

Zoals bijna iedereen het erover eens is, dat programma is mislukt, het was onmogelijk. Het probleem is een semantisch probleem. Wat de transducers afleveren zijn gewaarwordingen. De interne representaties van de machinetaal zelf zijn ongeïnterpreteerd; ze vormen een formeel systeem, maar hun vorm op zichzelf maakt niet dat ze iets betekenen, dat ze iets representeren. Alle betekenis die ze hebben moeten ze krijgen via het contact met de buitenwereld Het enige contact met de buitenwereld wordt gevormd door de transducers. De enige output die de transducers leveren is: "Er is stimulatie van die-en-die intensiteit op die-en-die plaats van het sensor-oppervlak". Op zichzelf is de aard van die stimulatie ook ongeïnterpreteerd, maar aangezien sensoren gevoelig zijn voor een beperkte range van stimuli, mogen we zeggen

dat de transducers gewaarwordingen afleveren als "Het is nu rood op die-en-die plaats van het sensor-oppervlak" of "Er is een lichtbalk met die-en-die orientatie op die-en-die plaats van het sensor-oppervlak" (Merk op dat ook in dit verhaal het probleem van de omgekeerde qualia kan optreden!) Dat is uitsluitend informatie over het eigen systeem! Waar we naar toe moeten zijn representaties van dingen in de buitenwereld, van de distale stimuli, niet van de proximale stimulatie We moeten een *vertaling* zien te krijgen van gewaarwordingsconcepten (of gewaarwordings- en motorconcepten) naar gewone concepten Dat betekent dat 'gewone' uitspraken over dingen en toestanden in de wereld, zoals "Dit ding is rood", equivalent moeten zijn met of definieerbaar in een aantal uitspraken die enkel verwijzen naar gewaarwordingen Maar het is eenvoudig niet waar dat onze gewone 'ding-uitspraken' equivalent zijn met pure gewaarwordingsuitspraken De uitspraak "Dit is rood" is alleen maar equivalent met de uitspraak "Roodheid is nu de gewaarwording" onder specificatie van de omstandigheden Onder een blauwe lamp, of bij een kleurenblinde, ontstaat geen ervaring van roodheid Die specificatie van de omstandigheden, zelfs al is dat alleen *'ceteris paribus'* of *'normaliter'*, verwijst zelf weer, behalve naar toestanden van het organisme, naar toestanden en dingen in de wereld De *vertaling* van gewaarwordingsuitspraken naar gewone uitspraken over de wereld is niet mogelijk (zie Chisholm 1957) Als de transducers alleen maar gewaarwordingsuitspraken afleveren is het probleem van de interpretatie van interne representaties niet opgelost

Het is ook mogelijk te zorgen dat de output van de transducers bij een robot bestaat uit volledige uitspraken in de ding-taal in plaats van in de gewaarwordingstaal Dit is een strategie die vaak wordt toegepast in de AI De transducers zelf worden dan niet gebouwd, maar vooronderstelld. Dit is het geval bij Winograd s SHRDLU Maar deze manoeuvre verplaatst het probleem alleen maar naar de transducers (44)

Nogmaals, het gaat hier om het probleem van de semantische interpretatie van interne representaties, waarbij we ervan uitgingen dat de interne representaties in de machinetaal uit ongeinterpreteerde labels bestonden plus geinterpreteerde 'logische regels Het gaat dus niet om het probleem hoe bepaalde combinaties van gewaarwordingen

een bepaald concept in een al geïnterpreteerde interne taal kunnen activeren. Dat proces vraagt niet om een equivalentie tussen gewaarwordingsuitspraken en ding-uitspraken, of om een logisch geldige afleiding van een concept uit gewaarwordingen. Het gaat erom of een ongeïnterpreteerde interne taal een volledige semantische interpretatie kan krijgen door toevoeging van de geïnterpreteerde output van de transducers. De transducers hebben tot taak de input van fysische stimuli om te zetten in een symbolische output. Die output representeert wat er op het sensor-oppervlak gebeurt, en is dus een gewaarwordingsuitspraak. Maar een ongeïnterpreteerd formeel systeem krijgt niet zijn semantische interpretatie uit enkel gewaarwordingsuitspraken en logische concepten, zoals de fenomenalisten met hun betekenistheorie tot hun spijt ontdekten. En wanneer je ontkent dat de transduceroutput uit gewaarwordingsuitspraken bestaat en zegt dat de transducers waarnemingsuitspraken afleveren - de output is per definitie symbolisch, een *uitspraak* in een interne taal - dan vooronderstel je dat de transducers voor de omzetting van fysische stimuli in symbolische output beschikken over een al geïnterpreteerde taal.

Fodor erkent het bestaan van dit soort problemen. Voor een deel noemt hij ze in zijn artikel over procedurele semantiek. Hij waarschuwt Tom Swift om een computer niet te zien als een semantische theorie (Fodor 1981a, 224). Toch wil hij aangeven hoe volgens hem zijn theorie uiteindelijk afgerond zou moeten worden om een oplossing voor het probleem van semantische interpretatie te vinden.

In een ongepubliceerd artikel 'Narrow content and meaning holism' geeft Fodor een mogelijkheid aan hoe het kan dat de interne representaties geïnterpreteerd zijn, inhoud hebben. Hij wil in dat artikel zijn theorie over een *language of thought*, een interne taal, verbinden met zijn theorie over de *modularity of mind* (zie 2.3.4). Hij spreekt van *'narrow content'* omdat hij, in aansluiting op zijn methodologisch solipsisme, een notie van inhoud nodig heeft die afziet van waarheidswaarden. In een reactie op de kritiek op zijn solipsisme-artikel zegt hij nog:

> "... if there *is* a narrow notion of content, it must be determined independent of interpretation: it must be

determined by functional role" (Fodor 1980a, 106)

Daar lijkt het erop of Fodor denkt dat *narrow content* ongeïnterpreteerd is, dus niet alleen geen referentie heeft maar ook geen betekenis. Het kan ook zijn dat hij, zoals elders, het vastleggen van de interpretatie gelijkstelt met het vastleggen van de referentie, en dan heeft *narrow content* wel een betekenis - opnieuw zorgt het niet goed onderscheiden van betekenis en referentie hier voor problemen (zie 4.5.1 en 4.5.2) Hoe het ook zij, in zijn *narrow content* - artikel is *narrow content* alleen onafhankelijk van waarheidswaarde en referentie, en dus wel geïnterpreteerd

Fodor stelt de zaken als volgt voor: wanneer je alleen kijkt naar de interne representaties en de transducer-output, dan heb je een ongeïnterpreteerd formeel systeem. Wanneer je ook kijkt naar de proximale stimuli die op de transducers inwerken, dan is de transducer-output geïnterpreteerd. Daarmee heb je alles wat nodig is voor het vastleggen van de *narrow content*. Kijk je nog verder, ook naar de distale bron van de proximale stimuli, naar de dingen in de wereld, dan heb je alles wat je nodig hebt voor het vastleggen van 'brede inhoud', een notie van inhoud die wel rekening houdt met waarheidswaarden en referentie. Voor een opake taxonomie van mentale toestanden, waarbij je mentale toestanden indeelt op hun inhoud, is een notie van 'nauwe inhoud' nodig. Twee mensen zijn dan in dezelfde toestand als ze beide geloven dat de morgenster rood is, en in verschillende toestanden als de een gelooft dat de morgenster rood is en de ander dat de avondster rood is. In het laatste geval heeft hun interne representatie een verschillende nauwe inhoud, maar dezelfde brede inhoud.

Dit voorstel van Fodor lijdt evenwel aan hetzelfde semantische probleem van het fenomenalisme dat hij aanvalt in de procedurele semantiek. De transducer-output gaat over gewaarwordingen, en niet over dingen. Om dit euvel te verhelpen stelt Fodor een drietraps cognitief systeem voor: transducers ontdekken de proximale stimuli en geven gewaarwordingsuitspraken af; 'modulen' (zie 2.3.4.) leiden uit de transducer-output de (zoals hij het noemt) fenomenologisch waarneembare eigenschappen van distale dingen af; en de centrale denkprocessen leiden uit die eigenschappen van dingen af wat men nog

meer weet over de wereld. Een fenomenologisch waarneembare eigenschap van bijvoorbeeld water is volgens Fodor natheid, of transparantheid; de chemische structuur van water is typisch *niet* fenomenologisch toegankelijk.

Fodor ontkent dat zijn drietrapsmodel fenomenalistisch is, omdat voorbij de modulen (in de richting van buiten naar binnen) alle niet-logische concepten uitgedrukt kunnen worden in termen van de fenomenologisch toegankelijke eigenschappen van dingen, en niet in termen van gewaarwordingsconcepten. Bovendien gaat het hem om de notie van nauwe inhoud, en hoeft hij geen onderscheid te maken tussen de mentale toestand van iemand die meent dat dat katachtige geelbruine dier een tijger is wanneer er inderdaad een tijger is, of iemand die hetzelfde meent wanneer er een gevlekte poema aankomt. Beide meningen hebben dezelfde nauwe inhoud, en veroorzaken (mede) hetzelfde gedrag. Zo ook hebben mijn meningen over water dezelfde nauwe inhoud als de meningen van mijn dubbelganger op Tweelingaarde. Ons water heeft dezelfde *fenomenologische* eigenschappen als het spul daar.

Maar de vertaling van gewaarwordingsuitspraken naar uitpraken over dingen moet nog steeds gemaakt worden! Het fenomenalisme zit hem niet in de derde trap van het cognitief systeem, niet in de denkprocessen, maar in de tweede trap, in de modulen. Die moeten gewaarwordingsuitspraken vertalen in uitspraken in termen van fenomenologisch waarneembare eigenschappen van dingen. Chisholm's redenering liet zien dat ook die vertaling onmogelijk is: "Dat ding heeft de (fenomenologisch waarneembare) eigenschap rood" is niet equivalent met "Er is een gewaarwording van rood". Let wel, het gaat hier nog steeds niet om een theorie van perceptie, maar om de vraag hoe de interne representaties aan hun inhoud komen. Gewaarwordingsuitspraken zijn alleen equivalent met uitspraken over de eigenschappen van dingen als je er 'normaliter' of 'ceteris paribus' bijzegt, en in dat 'normaliter' of 'ceteris paribus' wordt al een wereld van betekenis vooronderstelt.

De invoering van modulen die gewaarwordingsuitspraken omzetten in fenomenologische uitspraken kan op zichzelf de beschuldiging van fenomenalisme niet weerleggen, ook al zit het fenomenalisme dan niet meer in de centrale denkprocessen. De modulen moeten, om hun

(symbolische!) omzetting te kunnen maken, beschikken over een interne taal waarin op zijn minst het concept 'fysisch object' of 'ding' al geïnterpreteerd is, anders kan de omzetting van (fenomenalistische) eigenschappen van sensor-oppervlakken naar (fenomenologische) eigenschappen van *dingen* nooit gemaakt worden. Fodor zou nog kunnen tegenwerpen (hij doet dat zelf niet) dat de modulen niet echt een *vertaling* leveren. Ze zetten gewoon fenomenalistische eigenschappen om in fenomenologische, en dat is niet een logisch inferentieel, maar een puur causaal proces. De modulen hoeven dan niet te *weten* wat 'normaliter' inhoudt, ze *werken* normaliter op een bepaalde manier. Ik heb tegen zo'n tegenwerping twee bezwaren. Ten eerste vind ik het moeilijk in te zien wat een omzetting van de ene symbolische representatie in de andere in een interne taal anders kan zijn dan een vertaling. En ten tweede wordt de werking van de modulen voor wat betreft de semantische interpretatie van de interne taal *onbegrijpelijk*. Zoals Fodor zegt:

> "The nonlogical vocabulary of this language denotes (not proximal stimuli but) the phenomenologically accessible properties of distal objects" (Fodor n.c., 40).

Maar dat *stipuleert* hij zomaar; het probleem was juist hoe dat kan, als eerst het vocabulaire alleen over de proximale stimuli (gewaarwordingen) ging. Het fenomenalisme probeerde dat te verklaren, maar was niet houdbaar. Fodor zegt in zijn *narrow content* artikel alleen dat hij niet fenomenalistisch denkt, maar probeert niet te verklaren hoe je aan een geïnterpreteerd vocabulaire over eigenschappen van distale objecten komt (45) - hij probeert wel te verklaren hoe de interpretatie van de interne taal geheel vastgelegd kan worden als de termen voor logische concepten en die voor fenomenologisch waarneembare eigenschappen geïnterpreteerd zijn (46). Ik concludeer dat Fodor's *narrow content*-artikel geen oplossing geeft voor het probleem van de semantische interpretatie van interne representaties.

Fodor's *narrow content*-artikel is nooit gepubliceerd, hij verwijst er zelf nooit naar en lijkt er niet op door te gaan. Sinds 1984 richt hij zijn aandacht op de causale relaties tussen de wereld en interne representaties

Ofschoon Fodor in zijn solipsisme-artikel twijfelt aan de praktische mogelijkheid van een naturalistische psychologie, is hij in latere artikelen toch op zoek naar de relatie tussen interne representaties en de wereld. Volgens hem wordt de interpretatie van interne representaties vastgelegd door een of andere causale relatie tussen representatie en referent in de buitenwereld Hij betwijfelt de praktische mogelijkheid van een wetenschap die over die causale relaties gaat, maar wil toch aangeven wat de algemene structuur van die relaties is. Ik vraag me evenwel af of hij die mogelijkheid wel om de juiste redenen betwijfelt, en of het hier niet gaat om een *principiële,* mogelijkheid, veeleer dan om een *praktische* onmogelijkheid.

Volgens Fodor moet een naturalistische psychologie zich bezighouden met de relaties tussen organisme en buitenwereld. Deze psychologie moet generalisaties vinden van de vorm "A's uiting van 'water' verwijst naar water dan en slechts dan als A een causale relatie R heeft met ---". Het probleem is volgens Fodor wat er voor de --- moet worden ingevuld. 'Water' verwijst alleen naar water als het gaat over H_2O, als H_2O kan worden ingevuld voor ---. Maar in dat geval is het de chemie die zegt waar 'water' naar verwijst. We kunnen geen naturalistische psychologie van referentie bedrijven als we niet weten wat water *is;* wat de *wetmatige* karakterisering van de extensie van water is. Maar dat geldt voor alles waar we aan kunnen denken: we moeten eerst weten wat het *is,* en dat is de taak van de natuurwetenschap. We moeten dus wachten tot de fysica en de chemie compleet zijn, voor we met naturalistische psychologie kunnen beginnen. Vandaar de praktische onmogelijkheid van naturalistische psychologie Zo zegt Fodor over het behavioristische probleem van de fysische karakterisering van bijvoorbeeld potloden:

"If they really had to wait for the physicists to determine

> the description(s) under which pencils are law-instantiators, how would the psychology of pencils get off the ground?" (Fodor 1980a, 71)

Voor Fodor is het probleem van de fysische specificatie van de stimulus een *praktisch* probleem, een kwestie van wachten op de voltooiing van de fysica (47). De onmogelijkheid van een naturalistische psychologie is een *praktische* onmogelijkheid, omdat eerst de andere (natuur)wetenschappen compleet moeten zijn.

Maar de naturalistische psychologie zoals Fodor die ziet heeft *principiele* problemen. De naturalistische psychologie moet volgens Fodor het probleem oplossen hoe de semantische interpretatie van interne represetaties vastgelegd moet worden, en die interpretatie wordt volgens Fodor vastgelegd door de causale relaties van wereld naar representatie. Er blijft immers geen andere mogelijkheid over (zie ook Meijsing 1984, Tan 1983)?

We hebben al moeten verwerpen, vanwege het token-fysicalisme, dat een bepaalde fysische inscriptie noodzakelijk een bepaalde interpretatie heeft; mijn interne representatie van katten hoeft niet dezelfde te zijn als de jouwe, net zomin als mijn woord voor katten, 'kat', gelijk hoeft te zijn aan het jouwe (misschien 'cat' of 'Katze' of 'chat'). Het is een magische theorie te denken dat bepaalde symbolen uit zichzelf de namen van dingen zijn. Putnam zegt dat in zo'n geval de relatie tussen representatie en het gerepresenteerde een kwestie van 'metafysische lijm' is (Putnam 1981) Een gelijkenistheorie die stelt dat de relatie tussen representatie en het gerepresenteerde er een van gelijkenis is is ook niet houdbaar, omdat niet gespecificeerd kan worden in welk opzicht die gelijkenis bestaat (of, in geval van een neurale code, zelfs maar kan bestaan). En in 4.6.1 en 4.6.2 zijn argumenten gegeven tegen de procedurele semantiek en de functionele-rol-semantiek.

Over blijft dus de causale relatie als vastlegger van de semantische interpretatie. En omdat Fodor een fysicalist is, moet die causale relatie fysisch gespecificeerd worden, en mag er in die specificatie geen verwijzing meer voorkomen naar al geinterpreteerde representaties. Fodor's theorie zegt dat de interpretatie van een interne representatie datgene is wat de representatie veroorzaakt heeft. Zoals steeds is ook hier niet duidelijk of Fodor bedoelt dat de oorzaak de referentie

vastlegt of ook de betekenis.

Fodor baseert zijn schets voor een causale theorie van representatie op het werk van de filosoof Fred Dretske *Knowledge and the flow of information* (1981) Volgens Dretske wordt de semantische inhoud van een toestand in een systeem (b.v een bepaalde neurologische structuur) vastgelegd door de informatiestroom in dat systeem die die bepaalde toestand van het systeem veroorzaakt. Dretske geeft een verfijnde uitwerking van Fodor's intuitie dat er een causale keten is van distale stimulus naar proximale representatie (zie 4.5.1)

Een uitgebreide bespreking van Dretske's werk zou hier te ver voeren, voor de discussie van Fodor's theorie van semantische interpretatie kan ik met het volgende volstaan Dretske gaat uit van de klassieke informatietheorie van Shannon en Weaver en spreekt van een bron en een ontvanger Volgens hem heeft een bepaalde toestand (r) in de ontvanger de informationele inhoud dat de bron (s) de eigenschap F heeft dan en slechts dan als de waarschijnlijkheid dat de bron F is, gegeven die toestand van de ontvanger, 1 is (r heeft de informationele inhoud Fs desda p(Fs/r)=1) Dus een bepaalde interne representatie van mij heeft de informationele inhoud dat de lucht nu blauw is als in feite de lucht altijd blauw is als ik die representatie heb, en dat is het geval als mijn representatie veroorzaakt wordt door een blauwe lucht en door niets anders Maar mijn representatie heeft nog veel meer informationele inhoud Stel dat bijvoorbeeld altijd de zon schijnt als de lucht blauw is, dan geldt niet alleen p(blauwe lucht/representatie)=1 maar ook p(zon/blauwe lucht)=1, en dus p(zon/representatie)=1: de zon schijnt altijd als ik diezelfde representatie heb. Die representatie heeft dus ook de informationele inhoud dat de zon schijnt. Maar in Fodor's theorie heeft een interne representatie altijd maar een bepaalde semantische inhoud Volgens Dretske is de semantische inhoud van een bepaalde toestand (representatie) de meest specifieke informationele inhoud van die toestand Mijn bovengenoemde representatie r heeft wel meerdere informationele inhouden maar slechts een semantische inhoud, namelijk de inhoud dat de lucht blauw is, want dat is de meest specifieke informatie

Deze Fodor-Dretske theorie, dat de semantische inhoud van een interne representatie datgene is wat de representatie veroorzaakt, kent

echter verschillende problemen.

Het *eerste* probleem gaat over de specificatie van de oorzaken van interne representaties. Fodor eist dat die oorzaken wetmatig gespecificeerd moeten zijn. Maar hij *bedoelt* fysisch (of chemisch) gespecificeerd, natuurwetenschappelijk in ieder geval. Daarom kan hij geen genoegen nemen met een theorie die zegt dat potloden de oorzaak zijn van onze interne representaties van potloden (zie ook b.v. Field 1972). Hij wil een fysische beschrijving op grond waarvan potloden die bepaalde structuren in ons hoofd - representaties van potloden - kunnen veroorzaken. Alleen lijkt het wachten op een voltooide potloden-fysica hem te lang. Maar het is geen kwestie van wachten! Er zal nooit een potloden-fysica komen die verder is dan we nu zijn. We kunnen interne representaties hebben van oneindig veel waarvan geen fysische beschrijving bestaat. Wat is de fysische beschrijving van potloden, po's, poppen, om nog maar te zwijgen over positivisme, politiek en poëzie?

Fodor argumenteert in zijn artikel 'Special sciences' tegen het fysicalisme dat alles probeert te vertalen in fysische termen (zie 3.3.1). Dat maakt het vreemd dat hij nu voor zijn causale theorie van semantische inhoud toch een fysische beschrijving wil van de oorzaken van onze representaties. Ook zijn streven naar een theorie van 'hoe dingen eruit zien' (zie 'Narrow content' en Fodor 1984b), houdt in dat Fodor nu meent dat er in eerste instantie één soort beschrijving is voor 'de dingen'. Dit sluit aan bij zijn stelling in *The modularity of mind* dat observatie *niet* theoriegeladen is, en dat de input-systemen voor perceptie informationeel ingekapseld zijn (zie 2.3.4). Omdat 'de dingen' onze interne representaties veroorzaken in de zin van fysische veroorzaking, moet die éne soort beschrijving een fysische beschrijving zijn.

Fodor is in dit opzicht pijnlijk eerlijk: hij ziet het probleem van semantische interpretatie van interne representaties als hèt probleem van zijn theorie van het mentale, èn bij de oplossing van dat probleem mag er niet gesmokkeld worden. Een oplossing die ergens semanticiteit vooronderstelt, en niet in puur fysische termen gesteld is, is geen oplossing volgens hem. Dus een theorie die de veroorzakers van onze interne representaties indeelt anders dan op grond van hun (natuurwetenschappelijk) wetmatige eigenschappen, is geen oplossing.

Het *tweede* probleem van de Fodor-Dretske theorie heeft te maken met de oneindigheid van causale ketens, en is een parallel van het probleem voor een functionele-rol-semantiek dat te maken heeft met de oneindigheid van de keten van gevolgen (zie 4.6.2). In veel gevallen klopt Dretske's voorstel dat de meest specifieke informatie de semantische inhoud is. Als een representatie de informationele inhoud P heeft, heeft hij ook de informationele inhoud PvQ - immers p(PvQ/P)=1. De representatie heeft alles wat afgeleid kan worden uit P als informationele inhoud, maar alleen P als semantische inhoud Dat is intuïtief wel plausibel. Bij causale ketens zou dat betekenen dat de semantische inhoud van een representatie de meest proximale oorzaak is. Dat is echter heel vaak niet het geval. De meest proximale oorzaak van mijn representatie dat Napoleon de slag bij Waterloo verloren heeft is een zin in een boek over geschiedenis, en niet Napoleon's nederlaag. Maar de semantische inhoud van mijn representatie is niet dat een zin over Napoleon in een boek stond. Zodra mijn representaties niet door directe perceptie van de wereld, maar door sociale communicatie verkregen zijn gaat de theorie niet meer op; dan is de semantische inhoud van een representatie niet meer de meest specifieke informationele inhoud

Dretske heeft een oplossing voor dit probleem: de semantische inhoud van een bepaalde structuur is de informationele inhoud die niet meer ingebed is in een ruimere, meer specifieke informatie. Dus het lezen van de zin "Napoleon heeft de slag bij Waterloo verloren" veroorzaakt in eerste instantie inderdaad een interne structuur met de semantische inhoud dat die zin in het boek staat. Dat is de meest specifieke informationele inhoud, waarin de informatie over Napoleon is ingebed. Uit de eerste interne structuur, met als semantische inhoud de informatie over een bepaalde zin in een boek, ontstaat een tweede structuur, en *die* structuur heeft pas de semantische inhoud dat Napoleon de slag verloren heeft. Dat betekent dat de eerste interne structuur die rechtstreeks veroorzaakt wordt door de buitenwereld toch altijd de ruimste, meest specifieke informatie als semantische inhoud heeft. De meeste systemen komen niet verder dan die ruimste informatie. Dretske geeft het voorbeeld van een voltmeter die 7 volt aangeeft. *Wij* zien daaraan dat de spanning tussen de aansluitklemmen 7 volt bedraagt, in ons ontstaat op grond van de interne representatie

met de semantische inhoud dat de wijzer op 7 staat een representatie met als semantische inhoud dat de spanning 7 volt bedraagt Maar de voltmeter heeft geen enkele structuur met die semantische inhoud, al is die informatie ingebed in het feit dat zijn wijzer op 7 staat

> "Given the nature of such instruments, this information about the source is always embedded in larger informational shells depicting the states of those more proximal events on which the delivery of this information depends This, basically, is why such instruments are incapable of holding *beliefs* concerning the events about which they carry information" (Dretske 1981, 187)

Mensen kunnen volgens Dretske door die meer proximale gebeurtenissen heen de informatie over de distale bron oppikken, eenvoudige mechanismen kunnen dat niet Mensen 'kijken' door het informatiedragende medium heen naar de informatie van de distale bron, eenvoudige mechanismen pikken alleen de informatie over het medium op Dat is een algemeen erkende intuitie, en lijkt op de intuitie dat mensen de wereld (de distale bron) waarnemen, en niet alleen hun eigen gewaarwordingen (het medium) (zie b v Chisholm 1957)

Het probleem met dit verhaal van Dretske is evenwel, dat het in zijn theorie volstrekt duister blijft op grond van wat precies sommige mechanismen (mensen) *wel* naast interne structuren met als semantische inhoud de meest specifieke informatie over het medium, interne structuren krijgen met als semantische inhoud de ingebedde informatie over de distale bron, en andere mechanismen (voltmeters) *niet*. Op grond waarvan veroorzaken interne representaties van het medium van de informatie soms interne representaties van de bron en soms niet? Of is het afleiden van de ingebedde informatie iets wat mensen zelf *doen?* Hoe het ook zij, als vastlegger van de semantische interpretatie van interne representaties schiet deze theorie te kort Het is wel duidelijk wat de semantische inhoud is van de eerste door de wereld veroorzaakte representatie, dat is de toestand van het medium (een bepaalde zin staat in een boek, de wijzer van de voltmeter staat op 7) Maar wat is de semantische inhoud van de daardoor veroorzaakte

representatie? Dat Napoleon de slag verloren heeft, of de spanning tussen de aansluitklemmen 7 volt bedraagt? Of dat bepaalde toetsen van een typemachine zijn ingedrukt, of de bevestiging van de wijzer van de meter een bepaalde torsie ondervindt? Of misschien dat Blücher op het juiste tijdstip bij Waterloo arriveerde, of de weerstand in de keten waarin zich de voltmeter bevindt zoveel Ohm bedraagt? Al die informatie is ingebed in die eerste, meest specifieke informatie die de semantische inhoud vormt van de eerste interne representatie, maar welke vormt de semantische inhoud van de volgende interne representaties? Dretske's theorie kan alleen de semantische inhoud van de eerste interne representaties vastleggen.

Het *derde* en grootste probleem voor deze causale theorie van representatie is dat *misrepresentatie* volgens deze theorie onmogelijk is. Als de semantische inhoud van een bepaalde representatie zijn *oorzaak* is, dan moet die representatie die oorzaak representeren en niet iets anders Of als de semantische inhoud de meest specifieke informationele inhoud is, dan moet die inhoud wel die informatie bevatten Maar stel nu dat ik lees "De koning van Frankrijk is kaal" en dat dan ook geloof Die zin bevat helemaal niet de informatie dat de koning van Frankrijk kaal is - p(koning is kaal/mijn representatie)=0, want er bestaat helemaal geen koning van Frankrijk - en dus zou mijn mening niet die semantische inhoud kunnen hebben. De zin bevat misschien de informatie dat iemand iets over referentie wil uitleggen, maar dàt is niet de semantische inhoud van mijn mening (zie ook Cummins 1983) De semantische inhoud van geschreven of gesproken taal is vaak niet de informationele inhoud in de technische zin van Dretske, om de eenvoudige reden dat wat geschreven of gezegd wordt niet altijd waar is: een zin kan de zaken misrepresenteren. Een mening die gevormd wordt op grond van zo'n zin is dan ook een relatie tot een misrepresentatie. Maar volgens Dretske is de semantische inhoud van een representatie altijd een informationele inhoud, en misinformatie bestaat niet in de theorie. Immers, representatie r heeft *alleen* de informationele inhoud dat s F is als p(Fs/r)=1. Als s in feite niet F is kan de representatie nooit die (foutieve) inhoud hebben

Die onmogelijkheid van misinformatie levert ook problemen op in niet-talige situaties Zo is het fenomenalistische probleem uit 4.6.3 met deze theorie niet op te lossen Als mijn netvlies rood registreert dan draagt

deze proximale gebeurtenis niet de informationele inhoud dat er iets in de wereld (de bron s) rood is. Immers: p(dat ding is rood/rood is nu de gewaarwording)≠1. Ik kan een rode gewaarwording hebben door een wit voorwerp onder een (voor mij onzichtbare) rode lamp. Mijn interne representatie 'rood' zou dan de semantische inhoud 'rood of wit' hebben, want ze heeft de informationele inhoud dat iets rood is of wit (p(dat ding is rood of dat ding is wit/roodheid is nu de gewaarwording)=1). De theorie kan geen misrepresentatie toelaten.

Dretske probeert dit probleem van misrepresentatie op te lossen. Hij onderscheidt een leerperiode, waarin een bepaalde interne representatie, bijvoorbeeld 'rood', alleen veroorzaakt wordt door rode dingen, en de tijd daarna, waarin het kan voorkomen dat bijvoorbeeld een wit voorwerp de representatie 'rood' als causaal gevolg heeft. Deze oplossing is evenwel om twee redenen niet houdbaar. Ten eerste veronderstelt ze een principieel onderscheid tussen de periode waarin een bepaald concept geleerd wordt en de tijd daarna; maar er valt geen tijdstip aan te geven waarna ons gebruik van een concept ophoudt gevormd te worden en te veranderen en voortaan 'echt' operatief is. Ten tweede moet in deze oplossing gegarandeerd zijn dat in de leerperiode niet kan gebeuren wat daarna wel mogelijk is, namelijk dat een wit voorwerp de oorzaak is van de representatie 'rood'. Bij Dretske is dat gegarandeerd omdat in de leerperiode een leraar aanwezig is die de misrepresentaties corrigeert. Maar dat betekent dat wat de interpretatie van interne representaties vastlegt niet de causale relatie met de wereld is, maar het pedagogische optreden van de leraar (zie Fodor 1984a, 242). Dan wordt in de causale theorie van representatie de intentionaliteit van de leraar vooronderstelt; zijn interne representaties zijn al geïnterpreteerd. De theorie geeft niet een fysicalistische verklaring waarom interne representaties überhaupt iets representeren zonder een beroep te doen op intentionaliteit.

Fodor zoekt langs een andere weg een oplossing voor het probleem van misrepresentatie. Hij stelt dat een interne representatie datgene representeert wat zo'n representatie veroorzaakt *onder normale omstandigheden.* Hij gebruikt als voorbeeld de voltmeter, waarvan de wijzeruitslag onder normale omstandigheden de spanning tussen de aansluitklemmen representeert. Geeft de wijzer '0' aan, dan betekent

dat dat er geen spanning is. Maar alleen onder normale omstandigheden, want als de aansluitklemmen verroest zijn geeft de wijzer altijd '0' aan. Zo ook representeren onze interne representaties wat onder normale omstandigheden hun oorzaak is. Nu moet enkel nog worden aangegeven wat normale omstandigheden zijn. Volgens Fodor zijn onze intuïties wat normale omstandigheden zijn niet stabiel als we niet weten waar een representerend systeem voor dient. Hij geeft het voorbeeld van een munt in zijn zak, waarvan de doorsnede bij gelijke lichaamstemperatuur varieert met de temperatuur van de omringende ruimte, en bij gelijke omgevingstemperatuur varieert met zijn lichaamstemperatuur. Wat is de oorzaak die zorgt voor representatie en wat zijn de normale omstandigheden die bij een afwijking zorgen voor een misrepresentatie van de oorzaak? Bij een voltmeter is dat duidelijk: normale omstandigheden zijn die waarin de meter werkt zoals hij bedoeld is te werken. Er wordt een beroep gedaan op de teleologie, op de bedoelingen van de maker Normale omstandigheden voor mensen zijn volgens Fodor ook die waarin mensen functioneren zoals ze bedoeld zijn te functioneren. Hij doet een beroep op wat hij noemt *natuurlijke teleologie,* en zegt dat dat teleologie zonder intentionaliteit is.

Deze oplossing voor het probleem van misrepresentatie is echter al evenmin houdbaar. Hoe kom je erachter hoe wij zouden moeten functioneren? Je kunt niet goed stellen dat wat in de meeste gevallen gebeurt het normale geval is waarin we functioneren zoals we zouden moeten functioneren. Bijvoorbeeld, volgens de Bijbel is een mensenleeftijd 70 jaar, en soms 80 Maar eeuwenlang is de gemiddelde leeftijd van de mens niet meer dan 30 jaar geweest Die 70, 80 jaar lijkt ons meer normaal dan die 30, meer zoals we bedoeld zijn, en dat leek ook de schrijver van die bijbeltekst normaal. Maar tot voor kort hebben veruit de meeste mensen dat niet gehaald, en ook nu vinden we die hoge gemiddelde leeftijden alleen in de rijke landen. Toch zijn we geneigd al die vroege doodsoorzaken - ondervoeding, oorlog, natuurramp - te zien als een toevallige afbreking van het normale functioneren, de normale levensloop. We hebben wel een intuïtie over wat goed functioneren is, en wat toevallige afwijkingen van dat goede functioneren zijn, maar is die intuïtie hard te maken - fysicalistisch hard? Hoe kun je principieel onderscheiden tussen een toevallige afwijking in het functioneren van iemand - Moeder Natuur's ontwerp is

goed genoeg maar dit individu is helaas een 'maandag'-exemplaar - en een normale beperking inherent aan het menselijk systeem - Moeder Natuur's ontwerp is niet het best denkbare? Hoe kun je, bijvoorbeeld, onderscheiden tussen een oogafwijking en een beperking inherent aan de bouw van het oog? Functioneren we zoals we zouden moeten functioneren als we in leven blijven? Hoe lang in leven blijven? Lang genoeg om het voortbestaan van onze soort te garanderen? Bestaat onze soort al lang genoeg om te kunnen zeggen dat we functioneren zoals we zouden moeten functioneren, dat onze soort voortbestaat? Allemaal onbeantwoordbare vragen.

En tenslotte, als *vierde* probleem, wat gebeurt er in deze versie van de causale theorie van representatie met de intentionaliteit van propositionele attitudes? Het gedrag moest volgens de theorie van propositionele attitudes verklaard worden met een beroep op de inhoud van mentale representaties, op hoe de persoon de objecten van zijn meningen en wensen voor zichzelf representeert Maar volgens de causale theorie representeert een representatie in het normale geval gewoon zijn oorzaak. Het hele probleem van verschillende beschrijvingen en coreferentiële termen wordt opeens ontkend omdat de causale theorie moet aannemen dat er maar een beschrijving en één indeling van de wereld is (zie punt a). Zo kan het niet meer uitmaken of Oedipus denkt aan Iokaste of aan zijn moeder, omdat zijn interne representaties beide veroorzaakt zijn door dezelfde fysische entiteit, en dus hetzelfde moeten zijn volgens het principe 'gelijke oorzaken gelijke gevolgen'

Dretske probeert dit probleem onschadelijk te maken door te stellen dat interne representaties (=bepaalde interne structuren) selectief gevoelig zijn voor bepaalde aspecten van de bron. Hij zegt van zo'n structuur:

> "It is sensitive, *selectively sensitive* if you will, to that component of the incoming information that defines the structure's semantic content" (Dretske 1981, 180).

Dat zal best, maar dat betekent dat wat de semantische inhoud van een interne representatie vastlegt die component van de binnenkomende informatie of van de oorzaak is, waarvoor de representatie selectief

gevoelig is; dat wil zeggen dat wat de semantische inhoud vastlegt een eigenschap van de interne representatie is, namelijk die selectieve gevoeligheid, en dàt betekent dat die interne representatie, die interne structuur, toch al intrinsiek een bepaalde semantische inhoud ***heeft***, onafhankelijk van de binnenkomende informatie of van de oorzaak. De interne structuur heeft immers zelf dan al eigenschappen die hem gevoelig maken voor bepaalde aspecten, en die gevoeligheid bestaat al onafhankelijk van de binnenkomende informatie. Het is dus niet langer de (informationele) causale keten die de semantische inhoud vastlegt, maar de gevoeligheid van de interne structuur. Het is een intrinsieke eigenschap van de interne structuur die de semantische inhoud bepaalt. We zijn dan weer terug bij een magische theorie van interpretatie waarbij een bepaalde fysische structuur noodzakelijk een bepaalde interpretatie heeft; en die theorie hadden we juist verworpen (zie boven) ten gunste van een causale theorie. Dretske's manoeuvre van een selectieve gevoeligheid voor een bepaald aspect van de oorzaak brengt zijn causale theorie weer terug tot een magische theorie.

Mijn conclusie is dat Fodor's (en Dretske's) causale theorie van representaties misschien wel de referentie van interne representaties kan vastleggen - maar alleen in het normale geval! - maar in het geheel niet hun betekenis, zelfs niet in het normale geval En, zoals moge blijken uit het voorbeeld van Oedipus, het is niet de referentie van een interne representatie die de semantische interpretatie vastlegt: 'Iokaste' *betekent* niet 'mijn moeder'.

Zolang we leven in een volstrekt eenduidige wereld van fysische objecten, kan de causale theorie verklaren wat de semantische interpretatie van interne representaties is, namelijk hun oorzaak. Maar in zo'n wereld is de hele notie van interne representaties overbodig. Het is dan niet meer nodig om het gedrag te verklaren met een beroep op interne representaties, omdat in het normale geval één causale keten loopt van stimulus naar respons: de stimulus veroorzaakt de (inhoud van) de interne representatie, en de interne representatie veroorzaakt de respons. De interne representatie mist dan elke mediërende functie en kan als schakel in de causale keten gemist worden. In zo'n wereld zou het behaviorisme ook een bevredigende theorie zijn. Maar juist waar de notie van interne representatie nodig is, waar het, bij het verklaren van gedrag, erom gaat hoe het systeem

de wereld voor zichzelf representeert, schiet de causale theorie van representaties te kort, en moet ze de semantische interpretatie van interne representaties vooronderstellen, in plaats van ze te verklaren (zie ook Silvers 1985)

Een mogelijke oplossing voor dit probleem zou kunnen zijn om niet fysische *objecten* de oorzaak te laten zijn van interne representaties, maar *eigenschappen* van die objecten. De eigenschap van een bepaald object dat het zijn moeder is veroorzaakt in Oedipus een andere interne representatie dan de eigenschap dat het een aantrekkelijke vrouw, genaamd Iokaste, is. Deze mogelijkheid is evenwel niet open voor een fysicalistische theorie, want de benodigde eigenschappen zijn beslist geen *fysische* eigenschappen (zie ook punt a) Het zijn zelfs vaak überhaupt geen objectief bestaande eigenschappen *Beauty is in the eye of the beholder!* De semantische interpretatie van een interne representatie wordt vastgelegd door die eigenschap van de referent die in die representatie aan de referent wordt toegeschreven Dat wil zoveel zeggen als dat de semantische interpretatie van een interne representatie vastgelegd is als je weet wat de referentie is en wat de manier van presentatie van die referent oftewel de betekenis van de representatie is; dat wil zeggen, als je weet wat de semantische interpretatie is Waar, maar weinig verhelderend (47a)

Ofschoon Fodor er gelijk in heeft dat de semantische interpretatie van interne representaties niet alleen door hun functionele rol wordt vastgelegd, maar voor een belangrijk deel door de wereld, kan hij geen *fysicalistische* verklaring geven hoe de wereld de interpretatie vastlegt (48)

4.7. *Conclusie ten aanzien van Fodor's theorie van het mentale.*

Fodor is in moeilijkheden. Hij kan geen oplossing geven voor het probleem van de semantische interpretatie van interne representaties Hij probeert een oplossing te vinden voor het referentieprobleem en voor het betekenisprobleem - welke beide hij onvoldoende onderscheidt - en meent dat het probleem van de intentionaliteit van propositionele attitudes hetzelfde is als het probleem van de semantische interpretatie van de representaties. In deze paragraaf wil ik laten zien dat het

intentionaliteitsprobleem wel onafhankelijk van het semantische probleem bestaat (zie 4 5.3) en een centrale rol speelt in Fodor's moeilijkheden.

Met uiterste consequentheid heeft Fodor geprobeerd de grondslagen van een cognitiewetenschap te expliciteren, en een fysicalistische theorie van het mentale te geven. Ik denk dat zijn analyses en argumenten buitengewoon scherpzinnig en vaak juist zijn Voor hem geen vage formuleringen, geen onduidelijke aanduidingen hoe de cognitieve theorie van het mentale ongeveer er uit ziet Maar ik denk dat wat hij uiteindelijk gedaan heeft, de uitkomst van zijn werk, een soort *reductio ad absurdum* van de positie in kwestie is. Hij geeft uistekende redenen waarom de theorie op een bepaalde manier moet gaan· een gematigd fysicalisme, gecombineerd met een computationele-toestand-functionalisme, een theorie van propositionele attitudes die werkt met een interne taal, een theorie van mentale veroorzaking waarbij expliciete representaties een causale rol spelen bij het produceren van gedrag, een computationele theorie van mentale processen De eerlijkheid gebiedt hem daarbij te zeggen dat in dit plaatje van de cognitiewetenschap bepaalde onderwerpen niet aan bod kunnen komen· qualia zijn niet in te passen in het functionalisme, en er is geen praktische mogelijkheid voor een theorie van de semantische eigenschappen van interne representaties.

Nu is dat niet iets om luchthartig over te doen, dat zijn ernstige gebreken voor een theorie van het mentale. Maar wat Fodor niet gezien lijkt te hebben is dat zijn theorie misschien wel op deze manier *zou moeten* gaan, maar zo niet *kan* gaan.

Om een fysicalistische theorie van mentale veroorzaking te hebben moet Fodor stellen dat mentale representaties een fysische belichaming hebben Ze moeten, met andere woorden, expliciet gerealiseerd zijn in een organisme. En het is de vorm van zo'n representatie die een rol speelt in de veroorzaking van gedrag. Maar het valt niet in te zien waarom zo'n fysische structuur een *representatie van iets* zou zijn. Fysische structuren op zichzelf zijn niet geïnterpreteerd Ze spelen wel een functionele rol, maar we hebben gezien, in het voorbeeld van de schakende of onderhandelende computer, dat dat niet voldoende is om de interpretatie vast te leggen. En noch de theorie van procedurele semantiek, noch de causale theorie van representatie kan verklaren hoe de fysische structuren geïnterpreteerd zijn; ze moeten die interpretatie

al vooronderstellen

Ik denk dat Fodor's problemen voortkomen uit het feit dat hij twee incompatibele theorieen met elkaar wil combineren Enerzijds wil hij een theorie van het mentale die intentionaliteit, als kenmerk van het mentale, serieus neemt. Dat brengt hem, met een theorie van propositionele attitudes die opaak, in intensionele zinnen, toegeschreven moeten worden, op het postuleren van interne representaties in een interne taal Anderzijds wil hij een fysicalistische theorie van het mentale en een fysicalistische theorie van mentale veroorzaking. Dat brengt hem op zijn formaliteitsconditie, waarin hij postuleert dat interne representaties fysische (neurale) structuren zijn die hun causale rol spelen op grond van hun vorm

Maar dan duikt het probleem van de semantische interpretatie op, in de drie vormen die ik heb onderscheiden het referentieprobleem, het betekenisprobleem en het intentionaliteitsprobleem Het probleem ontstaat als volgt: ongetwijfeld hebben allerlei fysische structuren een semantische inhoud. ze verwijzen ergens naar, hebben een betekenis, zijn ergens op gericht. Zulke structuren worden gevormd door gesproken of geschreven taal. Maar die semantische inhoud is conventioneel, en had voor hetzelfde geld anders kunnen zijn. De relevante tekst om te citeren is hier.

> "When *I* use a word," Humpty Dumpty said in a rather scornful tone, "it means just what I choose it to mean - neither more nor less " "The question is," said Alice, "whether you *can* make words mean so many different things." "The question is," said Humpty Dumpty, "which is to be master - that's all " (Lewis Carroll, Through the looking glass).

Gebonden door conventies (sommigen wat minder gebonden dan andere, dat wil zeggen wat meer meester) kunnen mensen de taal gebruiken om hun bedoelingen duidelijk te maken Dat gebruik kan, afhankelijk van context en toehoorders, meer of minder gelukkig zijn om die bedoelingen over te brengen (zie b v Searle 1969) Maar het zijn mensen die uiteindelijk betekenis en referentie verlenen aan de uitdrukkingen die ze gebruiken (zie b.v Strawson 1963, 1971;

Donnellan 1971; Searle 1983). De semantische eigenschappen van gesproken en geschreven taal zijn afgeleid van de semantische eigenschappen van de propositionele attitudes van personen.

In Fodor's theorie zijn propositionele attitudes relaties tot interne representaties in een interne taal. De propositionele attitudes krijgen hun semantische eigenschappen van de semantische eigenschappen van de interne representaties (zie Fodor 1981a, 31 en 1984a, 247). Dus de interne representaties moeten hun semantische eigenschappen *absoluut, intrinsiek* hebben, en niet meer ergens anders vandaan.

Dat klinkt zeer plausibel, maar het gaat mis zodra men tevens wil beweren dat de interne representaties fysische structuren zijn. Van een interne taal kunnen we nog zeggen dat dàt de taal is die haar semantische eigenschappen absoluut of intrinsiek heeft. Zo wordt de interne taal gedefinieerd. Maar bij fysische structuren lukt dat niet; van fysische structuren *weten* we dat het niet het soort dingen zijn die intrinsiek semantische eigenschappen hebben. We kunnen wel een tijdlang stipuleren dat onze interne representaties (=fysische structuren) semantische eigenschappen zoals betekenis en referentie hebben, maar uiteindelijk zal *verklaard* moeten worden hoe ze die kunnen hebben.

Die verklaring waarom fysische interne representaties semantische eigenschappen hebben en wat de semantische interpretatie van de representaties vastlegt, lukt niet, en er is een reden waarom het niet lukt. Fodor veronachtzaamt namelijk het intentionaliteitsprobleem. Hij denkt dat, als hij maar aan kan tonen dat er een unieke semantische interpretatie van de interne representaties is, hij het intentionaliteits-probleem heeft opgelost, samen met het semantische probleem. Maar, zoals ik in 4.5.3 heb laten zien, dat is niet zo. Ook al is er slechts eén unieke interpretatie mogelijk, dan nog moet iemand die interpretatie *uitvoeren.* Searle's Chinese kamer-voorbeeld toont dat aan. Als niemand de interpretaties uitvoert blijven de interne structuren leeg, als zinnen in het Etruskisch voor ons. Ze blijven dan wel hun functionele rol spelen, maar het zijn geen interne representaties voor het systeem dat ze heeft. Het systeem begrijpt niets, is nergens op gericht, heeft geen intentionaliteit.

Alle problemen in Fodor's theorie wijzen in de richting van het intentionaliteitsprobleem. Laten we de zaken nog eens op een rijtje

zetten.

Fodor is een realist voor wat betreft propositionele attitudes de persoon heeft echt meningen en wensen enz. Voor een deel zijn die misschien onbewust, en misschien ook wel niet als entiteiten aanwezig Sommige meningen en wensen zijn misschien alleen maar aspecten van het gedrag, of disposities om iets te doen Maar andere soorten meningen en wensen hebben we in ieder geval echt Een onderscheid van Rorty (1970) kan hier verhelderend zijn Rorty maakt onderscheid tussen *mental features* en *mental occurrents*. *Mental features* zijn min of meer permanente mentale toestanden meningen, twijfels, preferenties, wensen, verwachtingen. Zij bepalen (mede) het gedrag, en zijn min of meer vaste kenmerken (althans binnen een niet te groot tijdsbestek) van een persoon De *mental occurrents* zijn gebeurtenissen, geen toestanden, en vaak overkomen ze de persoon. Voorbeelden hiervan zijn *raw feels,* kleursensaties, pijnscheuten, plotseling opkomende gedachten of beelden. Zoeven schoot door mij heen de prachtige aanvangsregel van Rilke's gedicht *Herbst:*
Herr, es ist Zeit. Der Sommer war sehr gross
Dat is een typisch voorbeeld van een *mental occurrent*. Zulke *mental occurrents* zijn niet behavioristisch te reduceren tot gedragsaspecten of -disposities (welk gedrag?), zulke *mental occurrents* hebben we in ieder geval echt, het zijn echte gebeurtenissen, echte episodes in ons

Fodor analyseert propositionele attitudes als relaties tot een zin in de interne taal, tot interne representaties Maar relaties tussen wie of wat en interne representaties? Fodor is hier opvallend vaag over. De nadruk ligt bij hem op de relatie en op de mentale representaties, en meestal zegt hij alleen maar dat propositionele attitudes relaties tot interne representaties zijn. Slechts een enkele keer zegt hij wie of wat de relatie heeft tot de representaties. Hij spreekt dan afwisselend van 'het organisme' (1975, 1981a, passim), van 'mensen' (1981a, 187), van 'wij' (1981a, 200) en van 'het subject' (1980a, 63). Aangezien Fodor het er een enkele keer over heeft dat ook sommige dieren propositionele attitudes kunnen hebben, kunnen we het er misschien op houden dat hijzelf het liefst zou willen zeggen dat het het (hele) organisme is dat in relatie staat tot zijn interne representaties Maar met 'organisme' bedoelt hij dan niet iets specifiek biologisch, het is meer een fysicalistische term die gebruikt wordt ter vermijding van

anti-fysicalistisch aandoende termen als 'subject' of 'persoon Maar hij bedoelt er ongeveer hetzelfde mee, namelijk de entiteit die de propositionele attitudes heeft Nogmaals, Fodor zelf gebruikt de termen door elkaar en geeft geen redenen voor zijn keuze. Het organisme heeft een propositionele attitude als het in een bepaalde relatie staat tot zijn interne representaties Het organism is dus nog iets anders dan de verzameling interne representaties met wat machinerie er omheen Om in relatie te kunnen staan met de representaties moet het organisme tegenover die representaties kunnen staan, de representaties zijn er voor het organisme

Fodor is voorstander van een theorie van mentale veroorzaking Aangezien *mental features* het gedrag (mede) veroorzaken moeten *mental features* er ook echt zijn, net als *mental occurrents*. Beide worden geanalyseerd als relaties tot interne representaties En aangezien volgens Fodor de veroorzaking een fysische veroorzaking moet zijn, moeten de interne representaties als fysische (neurale) entiteiten aanwezig zijn Het organisme staat dan in relatie tot fysische entiteiten wanneer het in een propositionele attitude is

In Fodor's computationele theorie zijn mentale processen transformaties van mentale representaties De representaties spelen een causale rol in de veroorzaking van het gedrag, en ze interacteren causaal met elkaar, op grond van hun vorm Die causale interrelaties lopen (tot op zekere hoogte) parallel met de inferentiele relaties tussen de inhouden van de representaties als er een logische relatie is tussen de inhouden P en Q, dan zal er vaak ook een causale relatie zijn tussen de representaties met die inhoud De mentale processen vormen dus in zekere zin een model van de logica. Dit is evenwel volgens Fodor *niet* omdat de interacties tussen de representaties nu eenmaal automatisch de wetten van de logica volgen - dan zouden mensen immers altijd volledig rationeel moeten zijn en dat zijn ze niet - maar omdat de wetten van de logica ook gerepresenteerd zijn en causaal bijdragen aan de veroorzaking van andere interne representaties Fodor zegt dat de postulaten van de logica gerepresenteerd zijn *door* het organisme, en dat ze expliciteren "what real reasoners know about valid inferences" (Fodor 1981a, 120) De planeten volgen de wetten van Newton, maar hebben ze niet intern gerepresenteerd en raadplegen ze niet De planeten volgen automatisch

de wetten van Newton Maar mensen hebben de wetten van de logica wel intern gerepresenteerd (zie voor kritiek op dit punt Stabler 1983, zie ook Matthews 1984, die beweert dat het op empirische gronden niet uit te maken valt wat expliciet gerepresenteerd is en wat niet) Maar dan wordt het onduidelijk wie of wat de logica toepast op de andere interne representaties Is dat misschien het hele organisme, dat wat het weet over logica intern gerepresenteerd heeft, en die kennis toepast op wat het denkt en meent (zijn andere representaties)? Het *hele organisme*, dat iets anders is dan de verzameling interne representaties, lijkt hier toch iets te *doen* met die interne representaties, als het niet zo is dat de verzameling van interne representaties automatisch volgens de wetten van de logica interacteert

Als het organisme (de mens) interne representaties *heeft*, kun je nog wel zeggen dat dat betekent dat die interne representaties intrinsiek semantisch geinterpreteerd zijn Dat is immers hoe je interne representaties *definieert*, als die entiteiten waar alle andere semantische eigenschappen van zijn afgeleid. Maar als het organisme in relatie staat tot interne representaties, en die representaties zijn fysische structuren, dan moet nog aangetoond worden hoe die fysische structuren een interpretatie krijgen Fodor slaagt er niet in te laten zien dat interne fysische structuren qua mentale representaties een unieke semantische interpretatie hebben Dat komt omdat dit probleem niet anders is dan het probleem om aan andere fysische inscripties een unieke semantische interpretatie toe te kennen (zie Putnam 1983). Interne representaties moeten net zo goed nog geinterpreteerd worden als andere tekens In het normale geval is zo'n unieke interpretatie van fysische inscripties wel te geven, en is er geen sprake van ambiguiteit of multi-interpreteerbaarheid Maar middelbare school leerlingen weten tot hun verdriet dat nog altijd gevraagd kan worden wat de auteur bedoelde met bepaalde inscripties, en receptie-esthetici weten dat gevraagd kan worden hoe de lezer een bepaalde inscriptie interpreteert (zie Meijsing 1981) Waar het om gaat is hoe de interne representaties geinterpreteerd worden door het organisme dat ermee in relatie staat, waarvoor het interne representaties zijn

Searle merkt op dat het niet uitmaakt of een bepaalde structuur in mij veroorzaakt is door iets in de buitenwereld, zolang ik dat niet weet

representeert die structuur voor mij niet dat iets in de buitenwereld (Searle 1980). En Fodor zegt iets wat daar veel op lijkt, wanneer hij spreekt over mentale beelden.

> "What makes my stick-figure an image of a tiger is not that it looks like one (my drawings of tigers don't look much like tigers either) but rather that it's *my* image, so I'm the one who gets to say what it's an image of. My images (and my drawings) connect with my intentions in a certain way; I *take* them as tiger-pictures .. " (Fodor 1975, 191).

Ik denk dat eenzelfde parallel bestaat tussen interne taal en natuurlijke taal als tussen mentale beelden en tekeningen. Het organisme moet de mentale beelden en de interne representaties nog interpreteren, nog *opvatten als* bepaalde beelden en representaties. Omdat Fodor interne representaties ziet als causaal werkzame fysische structuren - hij wil een mechanisme voor de relatie tussen organismen en proposities - maar tevens als representaties voor het organisme dat ze heeft, zit hij met een gigantisch probleem.

Zijn theorie van het mentale wil een fysicalistische theorie zijn. Hij wil een fysicalistische analyse geven van intentionaliteit, het kenmerk van het mentale. Hij geeft een analyse van een systeem met intentionaliteit, de persoon, en probeert dat uiteindelijk een fysicalistische analyse te laten zijn. Maar dat systeem heeft interne representaties in de vorm van fysische structuren *in* zijn hoofd. Door wie worden die structuren geïnterpreteerd, voor wie representeren ze iets? Door en voor de persoon (de persoon zelf?) wier intentionaliteit nog volledig nodig is. De intentionaliteit is niet een (semantische) eigenschap van de representaties, zoals Fodor denkt (en hoopt), maar blijft een eigenschap van de persoon (in Fodor's termen, van het hele organisme) De analyse van de persoon (van het hele organisme) maakt gebruik van een persoon met dezelfde eigenschappen. De persoon als systeem met intentionaliteit wordt niet verklaard maar voorondersteld.

In Moby Dick schrijft Melville over de getatoeeerde wilde, Queequeg, dat deze had

> "...written out on his body a complete theory of the heavens

> and the earth, and a mystical treatise on the art of attaining truth; so that Queequeg in his own proper person was a riddle to unfold; a wondrous work in one volume; but whose mysteries not even he himself could read ..."(49).

Volgens Fodor's theorie zijn we allemaal in een soortgelijke positie Weliswaar is bij ons niet onze huid, de buitenkant, getatoeeerd, maar zijn onze hersenen, de binnenkant "getatoeeerd". En die tekst is gedeeltelijk geïnterpreteerd door de functionele rol die de zinnen spelen. Maar wie of wat legt de volledige interpretatie vast, en voor wie representeert die tekst iets?

Fodor moet in zijn theorie van het mentale een volledige persoon, een systeem met intentionaliteit, vooronderstellen. Daarmee heeft hij geen fysicalistische oplossing voor het lichaam-geest probleem gegeven

5 DANIEL DENNETT EEN REDELIJK ALTERNATIEF?

5 1 Inleiding.

In zijn boek *The language of thought* begint Fodor zijn uiteenzetting en explicitering van de grondslagen van de cognitiewetenschap met een citaat van Lyndon B Johnson "I'm the only President you've got " Dat mocht misschien van toepassing lijken in 1975, nu is het zeker niet zo dat Fodor's versie van de cognitiewetenschap de enig mogelijke is Daniel Dennett biedt een andere, coherente, filosofische basis voor de cognitiewetenschap in zijn boek *Brainstorms* en vele artikelen (en ook al in *Content and consciousness* (1969))

We hadden in hoofdstuk 1 gezien dat de filosofie van de cognitiewetenschap in twee richtingen uiteenvalt men kan realist zijn voor wat betreft propositionele attitudes, of men kan instrumentalist zijn in dat opzicht We zagen dat Boden's oplossing voor het lichaam-geest probleem die keuze niet maakte, bij haar was het onduidelijk of in haar theorie bepaalde systemen echt meningen en wensen hebben, of dat het handig is om er meningen en wensen aan toe te schrijven. Beide opties zijn door respectievelijk Fodor en Dennett uitgewerkt Fodor is een uitgesproken realist voor wat betreft propositionele attitudes, Dennett is een instrumentalist

Dennett's instrumentalisme komt voort uit drie overwegingen, waarvan de eerste twee Ryleaans van oorsprong zijn. *Ten eerste* is hij ervan overtuigd dat

> 'even for creatures who are genuine selves, *there is nothing it is like* to believe that p, desire that q, and so forth" (Dennett 1978b, 32, zie ook 1 3.2).

Mensen hoeven nooit te merken dat ze bepaalde meningen en wensen hebben

Ten tweede vermeldt Dennett het het probleem dat ik het intentionaliteitsprobleem genoemd heb. Als er interne representaties bestaan in de vorm van fysische structuren in het hoofd, dan moeten dat representaties voor iemand zijn, iemand moet ze kunnen 'lezen' en

interpreteren. Dat kan volgen Dennett nooit de persoon zelf zijn, want die representaties zitten *aan de binnenkant* van de persoon, zij kan ze nooit *zien*. Het hebben van een mening kan dus volgens Dennett niet gelijk zijn aan het hebben van een interne representatie

Ten derde kunnen meningen en wensen alleen verklarend worden toegeschreven aan een systeem onder aanname van de rationaliteit van dat systeem. Maar niemand is volledig rationeel, dus de toeschrijvingen die op zo'n assumptie berusten kunnen niet letterlijk waar zijn.

In 5.2 zet ik de bovengenoemde overwegingen van Dennett uiteen Deze overwegingen brengen Dennett tot een instrumentalistische opvatting van propositionele attitudes In 5.3 schets ik Dennett's instrumentalistische voorstel: meningen, wensen, enz. kunnen volgens hem soms worden toegeschreven met predictief succes, maar systemen hebben ze niet echt

Dennett is net als Fodor uit op een fysicalistische theorie van het mentale, die intentionaliteit als kenmerk van het mentale op fysicalistische wijze wil verklaren Hij zit dan met het intentionaliteitsprobleem. hij moet enerzijds het bestaan van interne representaties aannemen, maar anderzijds vooronderstellen interne representaties een instantie die die representaties kan interpreteren, een instantie met intentionaliteit. Die instantie kan niet de persoon zelf zijn, maar moet aan de binnenkant, *in* de persoon zitten: een *homunculus*. In 5 4 geef ik Dennett's behandeling van dit probleem weer, alsook zijn oplossing ervoor, die samenhangt met zijn instrumentalistische voorstel. In 5.5 geef ik een uitgewerkt voorbeeld aan de hand waarvan Dennett zijn theorie van het mentale demonstreert.

Dennett's instrumentalistische, fysicalistische theorie van het mentale moet zowel laten zien dat een groot aantal vermeende mentale entiteiten (toestanden of eigenschappen of gebeurtenissen) niet bestaan, alsook hoe het komt dat we - tot nu toe - denken dat ze wel bestaan. In 5.6 bekritiseer ik Dennett's eliminatief materialisme, mede aan de hand van een voorbeeld. In 5 7 laat ik zien dat Dennett's theorie hem dwingt tot een volledige derde-persoonsvisie en tot verificationisme, ofschoon hij daar zelf ambigu tegenover lijkt te staan. En in 5.8 wil ik laten zien dat Dennett's fysicalistische theorie van het mentale toch op een of

twee plaatsen (afhankelijk van de mate van verificationisme) intentionaliteit vooronderstelt en dus onverklaard laat.

5.2. *Ryleaanse bezwaren tegen representaties.*

Zo'n vijfendertig jaar geleden viel Ryle, in *The concept of mind,* een theorie van het mentale aan die hij de intellectualistische mythe noemde. Dit is de opvatting dat gedrag veroorzaakt wordt door een scala van interne gebeurtenissen. het denken van privé-gedachten, het raadplegen van regels en recepten, het uitvoeren van redeneringen en berekeningen.

De cognitiewetenschap, zeker in de Fodoriaanse versie, lijkt heel veel op de positie die Ryle met zoveel verve belachelijk had gemaakt Cognitiewetenschappers spreken openlijk en zonder blikken of blozen over interne mentale representaties en over interne calculaties en computaties over die interne representaties. Ryle viel echter een positie aan die ook andere aspecten had. Veel van zijn aanvallen zijn gericht tegen een Cartesiaans dualisme en tegen geheimzinnige, paramechanische interne processen Maar de cognitiewetenschap is fysicalistisch, en de interne processen daar gepostuleerd zijn niet mysterieus paramechanisch maar gewoon mechanisch. De cognitieve wetenschap is heel wat plausibeler en verfijnder dan Ryle's stroman.

Maar Ryle viel nog iets anders aan ook. Hij was diep wantrouwig tegenover iedere theorie die interne representaties postuleert, omdat zo'n theorie leidt tot een oneindige regressie van homunculi voor wie die representaties representeren. De interne representaties zitten immers *in* het hoofd, waar de persoon zelf ze niet kan zien. Er moet dan dus ook in het hoofd een systeem zitten dat de representaties kan zien en interpreteren en ernaar handelen. Zo'n systeem moet, om dat te kunnen, zelf intentionaliteit hebben, een persoon zijn, een klein mensje, een homunculus. En wanneer we het gedrag van zo'n homunculus willen verklaren, heeft hij weer interne representaties nodig en een instantie in zijn hoofd om die representaties te lezen enz.

De cognitiewetenschap antwoordt hierop door te verwijzen naar de computer-metafoor. Mensen zijn net als computers, en computers zijn machines-zonder-geest die interne representaties manipuleren. Er is

geen sprake van een oneindige regressie: computers bestaan. We hebben echter in het vorige hoofdstuk gezien dat het niet zo eenvoudig ligt met de representationele theorie van het mentale. Op het eerste gezicht lijkt die zonder meer te combineren te zijn met een computationele theorie, waarbij problemen van homunculi zich niet voordoen: computaties gaan automatisch, en het is niet nodig dat de interne representaties van de machinetaal door iets of iemand geïnterpreteerd of begrepen worden, ze worden gewoon uitgevoerd. Maar bij nader inzien bleek dat de interpretatie van die representaties problematisch is. De interne representaties in de machinetaal zijn in zekere zin geïnterpreteerd, maar ze hebben niet de bedoelde interpretatie - ze gaan niet over de wereld, alleen over machinebewegingen. En zelfs al zou de representaties in de interne taal een unieke en juiste interpretatie toekomen, dan is er nog altijd iemand nodig die de interpretatie uitvoert, voor wie de representaties iets representeren. De theorie is bij nader inzien nog even gevoelig voor het homunculus-bezwaar als Ryle's intellectualistische mythe.

Dennett is een fysicalist, een functionalist, en een aanhanger van een computationele theorie van het mentale, maar hij neemt Ryle's bezwaar tegen interne representaties serieus. (Ik noemde dit in de vorige paragraaf zijn tweede overweging.) Zo zegt hij in een artikel uit 1982-1983, getiteld 'Styles of mental representation' over Fodor's computationele theorie van het mentale:

> "The idea, apparently, is that in order to have an effect, in order to throw its weight around, as it were, an item of information must weigh something, must have a physical embodiment, and what could that be but an explicit representation or expression of it? I suspect, on the contrary, that this is almost backwards. *Explicit* representations, by themselves (considered in isolation from the systems that can use them), may be admirably salient bits of the universe, off which to bounce photons or neurotransmitter molecules or marbles, but they are by themselves quite inert as information-bearers in the sense we need" (1982-83, 217).

Dennett wil wel spreken van representaties, maar hij wil een onderscheid maken tussen expliciete, impliciete en *tacit* representaties. Zijn gebruik van deze termen wijkt wat af van het gangbare (niet echt eenduidige) gebruik. Ik zal zijn definities geven (Dennett 1982-83).

Informatie is *expliciet* gerepresenteerd in een systeem als er, op een functioneel relevante plaats in het systeem, een fysisch gestructureerd object is dat gezien kan worden als instantie van een formule of een zin in een taal. Die taal heeft een syntaxis en een semantiek (het systeem moet dan ook een mechanisme hebben om die formule te lezen).

Een stuk informatie is *impliciet* gerepresenteerd, wanneer het logisch geïmpliceerd is door iets dat expliciet gerepresenteerd is. *Impliciet* hangt dus af van *expliciet*. Impliciet is niet helemaal hetzelfde als potentieel expliciet, omdat in ieder systeem veel meer impliciet gerepresenteerd is dan, vanwege tijdsbeperkingen alleen al, ooit expliciet kan worden.

In de zin van *tacit* die Dennett gebruikt is het precies andersom: *expliciet* hangt af van *tacit*. Ryle had gezien dat het postuleren van enkel expliciete representaties leidt tot een oneindige regressie, omdat die weer geïnterpreteerd moeten worden door een interne homunculus. Er moet uiteindelijk een systeem zijn dat directe *know how* heeft. Als men al kan zeggen dat het systeem die *know how* representeert, dan is dat niet expliciet en niet impliciet maar *tacit*. De *know how* moet in het systeem ingebouwd zijn op een manier die niet vereist dat hij expliciet in het systeem gerepresenteerd wordt.

Een voorbeeld: neem een zakrekenmachientje (niet een heel moderne, die werken weer heel anders). Daar zijn geen rekenkundige proposities in code in te vinden. De enige expliciete representatie van cijfers is op het eerste gezicht op de knopjes, en, bij output, in het verlichte raampje. Maar er is nog meer expliciete representatie. Bij vermenigvuldiging, bijvoorbeeld 6x7=?, telt de calculator snel (in binaire notatie) 7+7+7+7+7+7 op, en bewaart de tussenuitkomsten in zijn buffer. Dat zijn andere uitkomsten dan wanneer hij 7x6=? uitvoert. In het eerste geval zijn de tussenuitkomsten14, 21, 28, 35 en in het tweede geval 12, 18, 24, 30, 36. Dat verschil is te zien. Ook hier is sprake van expliciete representatie van getallen, maar waar representeert de calculator de ware rekenkundige proposities? Het ding is zo gemaakt dat het die regels volgt, zonder ze ooit te hoeven

raadplegen. Ze zijn alleen *tacit* gerepresenteerd. Niet expliciet, want ze staan nergens, en ook niet impliciet, want ze volgen uit geen expliciet gerepresenteerde informatie

De onderscheidingen die Dennett hier maakt zijn nog niet vermeld in zijn *Brainstorms*, maar de theorie van het mentale die daar ontvouwd wordt gebruikt die onderscheidingen wel (impliciet denk ik, en niet *tacit*).

Met deze onderscheidingen kan Dennett aan de Ryleaanse bezwaren ontkomen zonder te hoeven zeggen dat er helemaal geen interne representaties bestaan. De meeste representaties, die in Fodor's theorie het gedrag (mede) veroorzaken, zijn volgens Dennett *tacit*. Dat betekent voor hem dat het gedrag zodanig is dat het handig is om die representaties instrumentalistisch toe te schrijven. Maar ze zijn nergens expliciet aanwezig, en zijn dus ook geen echte oorzaken. Als instrumentalist kan Dennett niet meer zeggen dat propositionele attitudes het gedrag veroorzaken, het zijn volgens hem aspecten van het gedrag, of handige ficties om het gedrag te systematiseren. Dit strookt ook met zijn eerste overweging, namelijk dat het *not like anything* is om zogenaamde meningen en wensen te hebben Het zijn enkel aspecten van het gedrag.

Waar wel sprake is van expliciete interne representaties, zoals de tussenuitkomsten in het rekenmachientje, hebben die representaties niets te maken met propositionele attitudes Zulke representaties moeten er wel zijn - geen computatie zonder representatie, er moeten symbolen zijn om gemanipuleerd te worden - maar dat zijn geen representaties *voor de persoon*. Nogmaals, de persoon kan die interne representaties niet zien, ook al sturen ze het gedrag.

Naast de appreciatie van de Ryleaanse bezwaren tegen representaties heeft Dennett misschien nog een andere reden, zijn derde overweging, voor zijn instrumentalisme. Het toeschrijven van meningen en wensen op grond van het gedrag van een systeem vooronderstelt rationaliteit bij dat systeem. Alleen van een rationeel systeem zijn mening-toeschrijvingen voorspellend. Maar niemand is volledig rationeel. Een mening-toeschrijving voorspelt wat iemand zou moeten doen, maar vaak niet wat hij in feite doet. Dus zijn mening-toeschrijvingen louter heuristisch.

Fodor (1981a, 1985) ziet deze redenering als Dennett's voornaamste

reden voor zijn instrumentalisme Ik vind evenwel Dennett's redenering op dit punt allerminst helder Men kan hem zo lezen als Fodor doet, alsof de contrafactische assumptie van rationaliteit zou impliceren dat meningen niet echt bestaan Maar Dennett zegt ook dat alle evolutionair succesvolle systemen "... in virtue of their rationality ... can be supposed to share our beliefs in logical truths" (Dennett 1978b, 9). Hier vindt Dennett niet alleen dat evolutionair succesvolle systemen rationeel zijn, maar ook dat ze meningen hebben.

Ik ben geneigd deze door Fodor gesuggereerde reden van Dennett voor zijn instrumentalisme als apocrief te beschouwen. Volgens mij tellen alleen zijn Ryleaanse overwegingen.

5.3. Intentionele systemen

In een artikel uit 1971, getiteld 'Intentional systems' (herdrukt in Dennett 1978b) legt Dennett de grondslag voor zijn theorie met het concept van een systeem waarvan het gedrag verklaard en voorspeld kan worden - ten minste bij tijd en wijle - door het systeem meningen en verlangens en verwachtingen en bedoelingen en vermoedens en twijfels enz toe te schrijven Zulke systemen noemt hij intentionele systemen, en zulke verklaringen en voorspellingen intentionele verklaringen en voorspellingen, vanwege de intensionaliteit (50) van het taalgebruik van mening en verlangen (en verwachting, bedoeling, vermoeden en twijfel enz.).

Dit concept van een intentioneel systeem moet Dennett's hele theorie van het mentale schragen, en een brug leggen tussen het intentionele domein van onze 'common-sense' wereld van personen en handelingen en de 'common-sense' psychologie, en het niet-intentionele domein van de fysische wetenschappen

> "That is a lot to expect of one concept, but nothing less than Brentano himself expected when, in a day of less fragmented science, he proposed intentionality as the mark that sunders the universe in the most fundamental way· dividing the mental from the physical" (Dennett 1978b, 22).

Allereerst dient erop gewezen te worden dat een bepaald ding alleen een intentioneel systeem is in relatie tot de strategieën van iemand die het gedrag van dat ding probeert te verklaren en voorspellen. Dennett geeft ter verduidelijking een voorbeeld (Dennett 1978b, 4-9).

Beschouw het geval van een schaakcomputer en de verschillende strategieën die men als zijn tegenspeler kan innemen om zijn zetten te voorspellen. Er zijn drie verschillende houdingen die men tegenover de schaakcomputer kan innemen.

Allereerst is er de *ontwerphouding*. Wanneer men precies weet hoe de machine en het programma zijn ontworpen, dan kan men het bedoelde antwoord op iedere zet voorspellen door de instructies van het programma op te volgen. Die voorspellingen zullen uitkomen zolang de computer zich gedraagt zoals hij zich moet gedragen - zonder storingen in de machine of *bugs* (programmeerfouten) in het programma. Al dit soort voorspellingen berusten op de notie van *functie;* de machine heeft een bepaalde functie, hij is bedoeld om te schaken, en alle programma-onderdelen hebben weer hun functie in dat geheel, een eigen taak, zoals bijvoorbeeld het genereren van legale zetten, het evalueren van zetten, het bijhouden van het geheugen enz.

De ontwerphouding nemen we meestal in tegenover mechanische voorwerpen, bijvoorbeeld:"Als de benzine opraakt gaat er een lichtje branden" of "Het verwarmingselement slaat af als de strijkbout op de ingestelde temperatuur is". We maken voorspellingen op grond van kennis van het functionele ontwerp, ongeacht de fysische constitutie van het systeem in kwestie.

Ten tweede is er de *fysische houding*. In deze houding zijn onze voorspellingen gebaseerd op de werkelijke fysische toestand van het voorwerp in kwestie, en we passen onze kennis van de natuurwetten erop toe. Alleen vanuit de fysische houding kunnen we verklaren waarom bepaalde systemen het niet doen, niet goed functioneren; bijvoorbeeld: "De auto start niet want de bougies zijn nat" of "Als je de stekker niet in het stopcontact steekt wordt de strijkbout niet heet".

Men neemt zelden tegenover een computer de fysische houding aan, ofschoon een onderhoudsmonteur dat wel eens doet. Maar om voor de zetten van de schaakcomputer voorspellingen te doen vanuit de fysische houding zou gekkenwerk zijn. In principe is het te doen: men

kan de effecten van de input-energieen volgen door de machine heen tot uiteindelijk letters tegen papier worden gedrukt en het antwoord verschijnt Maar de hoeveelheid tijd en moeite die dat zou kosten zou ongelofelijk groot zijn

De beste schaakcomputers zijn tegenwoordig niet meer voorspelbaar vanuit de fysische of vanuit de ontwerphouding Zelfs voor hun eigen ontwerpers zijn ze te complex geworden om de ontwerphouding aan te nemen Programmeurs kunnen door hun eigen programma verslagen worden

De beste methode om te winnen is dan ook de zetten van de computer te voorspellen door te bedenken wat de beste of meest rationele zet zou zijn, gegeven de regels en doeleinden van het schaakspel Men neemt dus niet alleen aan dat de machine werkt zoals hij ontworpen is, maar ook dat het ontwerp optimaal is, dat de machine de meest rationele zet zal 'kiezen' Men doet alsof de machine rationeel is, relatief ten opzichte van een doel of een hierarchie van doelstellingen, en een bepaalde hoeveelheid informatie De vraag die men bij de voorspelling stelt is wat is de meest rationele zet voor de computer, gegeven doelstellingen x, y, z, , voorwaarden a, b, c, . (de schaakregels) en informatie (ook eventuele foutieve) over de huidige stand van zaken p, q, r, . .? Deze houding, met zijn aanname van rationaliteit, is de *intentionele houding*, de voorspellingen die men doet zijn intentionele voorspellingen, men beschouwt de computer als een intentioneel systeem

Men kan de informatie die de computer bezit meningen noemen, zijn doeleinden verlangens, de notie van het hebben van informatie is even intentioneel als die van het hebben van een mening, en een doel even intentioneel als een verlangen Het gaat er niet om of de schaakcomputer *echt* meningen en verlangens heeft, de definitie van een intentioneel systeem zegt niet dat zulke systemen *echt* meningen en verlangens hebben, maar dat men hun gedrag het beste kan verklaren door meningen en verlangens eraan *toe te schrijven*. De beslissing om een intentionele houding aan te nemen is volstrekt pragmatisch, en is niet intrinsiek juist of onjuist

Wanneer het gedrag van een systeem al te irrationeel wordt, gezien zijn informatie en doeleinden, worden we gedwongen af te stappen van onze intentionele houding, die immers rationaliteit veronderstelt, en

gaan we over op de ontwerphouding Dit is de fundamenteel andere houding die we soms aannemen tegenover psychiatrische patienten Er wordt psychiatrische verplegers vaak op het hart gedrukt zich niet te laten 'meezuigen' door de patiënten De meningen en wensen die de patiënt uit worden in de therapie niet gezien als verhelderend voor voorspelling of verklaring van het gedrag, maar als, vaak zeer overtuigende, rationalisaties De therapie kan pas slagen als men dóór die zogenaamde meningen en wensen heen kan zien dat de patiënt door vroegere ervaringen als het ware is geprogrammeerd, *ontworpen*, om zich op een bepaalde manier te gedragen. Dennett noemt als voorbeeld dat de manier waarop een verpleger een obsessief contrasuggestieve patient manipuleert iets heel anders is dan normale interpersoonlijke interactie

Het is niet nodig dat de toegeschreven meningen en doeleinden ergens in het systeem expliciet gerepresenteerd zijn. Dennett (1982, 107) noemt als voorbeeld een kritiek op een bepaald schaakprogramma· "Het meent dat het zijn koningin vroeg in het spel moet brengen". Hier wordt een mening toegeschreven op een nuttige en voorspellende manier, want bij dat programma kun je er meestal op rekenen dat je de koningin het bord rond kunt jagen. Maar nergens in het programma is een expliciete instantiatie van een formule te vinden als: "Ik moet mijn koningin vroeg in het spel brengen". Het programma is alleen zo ontworpen dat het dat doet, die mening is *tacit* gerepresenteerd.

Het concept van een intentioneel systeem is ontologisch neutraal. Het abstraheert van vragen over de samenstelling, constitutie, bewustzijn enz van de entiteiten die eronder vallen.

Maar Dennett is wel een fysicalist. Hij is, om redenen die ik boven bij Fodor beschreef, geen type-fysicalist. Hij is, eveneens om dezelfde redenen als voor Fodor golden, ook geen Turingmachine-functionalist Dennett heeft een aantal nieuwe namen bedacht om zijn positie te karakteriseren Naast type-fysicalisme en token-fysicalisme is er ook *type-functionalisme, token-functionalisme* en *type-intentionalisme.* Al deze -ismen geven antwoord op twee vragen. 1) "Wat zijn mentale gebeurtenissen en toestanden?" en 2) "Wat hebben twee mensen gemeen als ze een mentale toestand of gebeurtenis gemeen hebben, b v beiden menen dat sneeuw wit is?" Alle genoemde -ismen zijn vormen van fysicalisme, en geven op vraag 1 het antwoord "Alle mentale

toestanden en gebeurtenissen zijn fysische toestanden en gebeurtenissen" Het type-fysicalisme antwoordt bovendien op vraag 2:

(x) (x meent dat sneeuw wit is = Px)

waarbij P een fysisch predikaat is. Het token-fysicalisme ontkent deze type-identiteit.

Alle vormen van functionalisme stellen dat mentale toestanden en gebeurtenissen *functionele* fysische toestanden en gebeurtenissen zijn. Het Turingmachine-functionalisme, een variant van het type-functionalisme (maar een token-fysicalisme) antwoordt op vraag 2·

(x) (x meent dat sneeuw wit is = x realiseert een Turingmachine k in logische toestand A)

Het token-functionalisme ontkent deze type-identiteit. Het type-intentionalisme, een soort token-functionalisme, antwoordt op vraag 2:

(x) (x meent dat sneeuw wit is = aan x kan met voorspellende waarde de mening worden toegeschreven dat sneeuw wit is).

Hier komt de notie van een intentioneel systeem om de hoek kijken. Wanneer het predictieve waarde heeft aan een systeem meningen (en wensen) toe te kennen dan is dat systeem een intentioneel systeem. Die notie is ontologisch neutraal, maar daarom juist compatibel met het fysicalisme

Dennett geeft aan deze versie van het token-functionalisme de naam 'type-intentionalisme' iedere mentale gebeurtenis is de een of andere fysische, functionele gebeurtenis, en de typen van mentale gebeurtenissen worden niet bepaald in een reductionistische taal maar door een ordening van de termen die we in het alledaagse taalgebruik gebruiken - we verklaren *wat meningen zijn* door de notie van een meningen-hebbend systeem te systematiseren.

Nu is Dennett ook geen type-intentionalist over de gehele linie. Hij gelooft namelijk niet dat onze alledaagse manier om mentale kenmerken en entiteiten aan te duiden in alle gevallen echte kenmerken en entiteiten aanduidt. van ons bekende mentalistische idioom verwijst nergens naar omdat het conceptueel ongelukkig en incoherent is. Niet alleen zijn *meningen* en *pijnen* geen goede theoretische *dingen* (zoals electronen of neuronen), maar de *toestand van menen-dat-p* is niet een definieerbare theoretische *toestand*, en het *attribuut pijn-hebben* is niet een fatsoenlijk theoretisch *attribuut*.

Ten aanzien van sommige (maar niet deze) mentale entiteiten is Dennett een type-intentionalist, ten aanzien van andere zogenaamde mentale entiteiten is hij een eliminatieve materialist (Dennett 1978b, XIV-XX).

5.4. Homunculi op het subpersoonlijke niveau.

"What makes the clown's clowning clever?" vroeg Ryle zich meer dan dertig jaar geleden met alliteratieve wellust af. En de leer die Ryle aanviel zegt het volgende: Wat de kunsten knap maakt is dat ze het gevolg zijn van een aantal mentale operaties (computaties, berekeningen) binnen in de clown. Dat kan volgens Ryle niet juist zijn want de kunsten kunnen enkel knap zijn als de mentale computaties knap zijn en wat maakt de mentale computaties knap? Wat volgens hem de kunsten knap maakt zijn kenmerken van de kunsten, en niet van de oorzaken van de kunsten want dat leidt tot een oneindige regressie. De clown kan niet de codes van zijn kunst gerepresenteerd hebben op de wanden van zijn geest, ergens *in* zijn lichaam, want wie moet dan de ogen hebben om die codes te lezen? Een knappe homunculus *in* de knappe clown.

Nee, volgens Ryle is er een publiek toneel, een persoonlijk niveau waar alles zich afspeelt. de kunsten van de clown, en de aspecten van die kunsten die ze knap maken, de regels van de kunst van het clownen, voor zover die bekend zijn; ook berekeningen en calculaties spelen zich af op dat publieke toneel, op schoolborden en schriften. Binnenin personen, op het subpersoonlijke niveau, *is* geen prive-toneel, waar zich andere kunsten afspelen, waar wordt gerekend en waar regels worden gevolgd. Binnenin personen is geen voer voor psychologen, alleen voer voor fysiologen.

Er is volgens Ryle geen expliciete interne representatie, wel een fysiologisch systeem dat zonder meer een zekere *know how* heeft. Als je daar al van representatie wilt spreken - wat Ryle niet doet - dan is dat *tacit* representatie in de zin van Dennett. Ryle zou spreken van gedragsdisposities. Natuurlijk erkent Ryle wel het bestaan van een aantal bekende verschijnselen zoals hoofdrekenen, *sotto voce* repetities van een toespraak, *silent soliloquies* zonder de lippen te bewegen enz.

Maar dat zijn hooguit mentale operaties op perifere expliciete representaties, aan de buitenste randen van de persoon. Verder, of dieper, naar binnen zijn er volgens Ryle geen expliciete interne representaties en geen computaties.

Helaas kan dit verhaal niet kloppen. Het valt eenvoudigweg niet te ontkennen dat er binnenin een persoon allerlei computaties plaatsvinden. In onze visuele perceptie is allerlei informatie over de minimale verschillen tussen beide netvliesbeelden, over de toestand van ons vestibulair systeem, over saccadische oogbeweging, *ingecalculeerd*. Er wordt daarbinnen heel wat afgerekend, en die berekeningen zijn transformaties van representaties. Er moeten dus wel expliciete interne representaties zijn.

Fodor neemt precies de tegenovergestelde positie in als Ryle. Volgens hem is er wel degelijk 'binnenin' de persoon van alles gaande, computaties en calculaties. En, zegt hij, "no computation without representation" (Fodor 1975, 34). Zijn representaties zijn causaal werkzaam, en daarom expliciet. Hij is niet bang voor homunculi omdat in een computer de representaties zichzelf begrijpen.

Helaas kan ook dit verhaal niet kloppen. Met zijn analyse van de procedurele semantiek heeft Fodor het onschadelijk maken van homunculi op losse schroeven gezet: computers hebben geen last van een oneindige regressie van homunculi. Het is in een computer niet nodig dat er steeds een instantie is die de interne representaties begrijpt: op het niveau van de machinetaal worden de instructies niet meer geïnterpreteerd maar direct uitgevoerd. Dat komt omdat de machine zo gebouwd is dat de causale werking van de structuren in de machinetaal correspondeert met de betekenis ervan. In die zin is de machinetaal geïnterpreteerd. Vergelijk: ik sluip achter een lezend persoon en schreeuw plotseling: "Schrik!"; de persoon schrikt, en de causale werking van mijn bevel correspondeert met de inhoud ervan (Dennett 1978b, 247). Maar de interpretatie van de machinetaal gaat alleen over machinetoestanden en is niet de bedoelde interpretatie van de interne representaties. En we zagen in 4.5.3 dat zelfs als er een unieke, bedoelde interpretatie van de interne taal zou zijn, er nog altijd iemand moet zijn voor wie de representaties die interpretatie hebben.

Dennett's positie houdt zo'n beetje het midden tussen die van Ryle

en van Fodor. Zijn theorie is een soort synthese van wat hij in beiden goed vindt. Hij neemt, zoals gezegd, Ryle's homunculus-argument serieus, en lijkt te zien dat Fodor met zijn expliciete representaties op dit punt in problemen is (51). Maar hij is het niet met Ryle eens wanneer deze de vraag: "Wat maakt de kunsten van de clown knap?" wil beantwoorden met een conceptuele analyse. Dàt is volgens Dennett geen psychologie. De psychologie moet wel degelijk de vraag beantwoorden: "what makes for intelligence?" (Dennett 1978b, 12).

Wat is er mis met een conceptueel antwoord? Volgens Dennett is dat in zekere zin circulair. Hij laat zien dat Ryle geen fysicalistische verklaring van gedrag kan geven. Ryle probeert niet, zoals Skinner, mentalistische termen uit te drukken in termen van stimulus- en respons-variabelen. Zijn explicaties zitten vol intentioneel taalgebruik. Zijn uitleg van ijdelheid (om een ander voorbeeld te nemen) is niet dat het een dispositie is om bepaalde bewegingen of geluiden te produceren, maar een dispositie om te proberen op te vallen, kritiek te negeren, over zichzelf te praten, de herinnering aan blunders te vermijden enz. (Ryle 1949, 86).

Als we spreken op het niveau van het hele organisme, op het niveau van de persoon, spreken we nu eenmaal in intentionele termen. Dat geldt zowel als we het over mensen hebben, als wanneer we over dieren spreken. Als een muis in een *T-maze* links een kat bij de uitgang kan zien en rechts kaas, dan voorspellen we dat ze rechtsaf gaat, omdat ze meent dat links een kat zit en rechts kaas, en liever wil eten dan gegeten worden. Mensen en dieren zijn intentionele systemen; we praten er in intentionele termen over. Dat doet Ryle ook; de kunsten van de clown zijn knap omdat ze (onder andere) onverwacht zijn. Maar, zegt Dennett,

> "In the end, we want to be able to explain the intelligence of man, or beast, in terms of his design, and this in turn in terms of the natural selection of this design; so whenever we stop in our explanations at the intentional level we have left over an unexplained instance of intelligence or rationality" (Dennett 1978b, 12). "Intentional theory is vacuous as psychology because it presupposes and does not explain rationality or intelligence" (Dennett 1978b, 15)

In tegenstelling tot Ryle, die wel met zijn analyses stopt op het intentionele niveau, heeft Skinner zoiets (52) gezien. Zijn reactie was om te proberen voorspellingen puur in niet-intensioneel taalgebruik te stellen, door lichamelijke responsen op fysische stimuli te voorspellen. Dennett laat zien dat dit alleen schijnbaar gelukt is. Stel dat een muis in een Skinner-box getraind is om precies vier stappen vooruit te doen en met haar neus een pedaal in te drukken. Als we het pedaal een paar centimeter zouden verschuiven, zou Skinner voorspellen dat de muis met haar neus in de lucht zou duwen, en geen vijfde stap zou maken. Het lukt ook niet om een respons te beschrijven als een bereikt effect in de omgeving: het pedaal gaat omlaag. Want stel dat de muis getraind is om het pedaal onder de langste van twee lichtbuizen in te drukken. Ze zou dan immuun moeten zijn voor de Müller-Lyer illusie, want de respons was, *ex hypothesi*, het indrukken van het pedaal onder de lamp die langer *is*, niet die langer *lijkt*. Maar muizen, duiven en vissen zijn gevoelig voor visuele illusies van lengte (Dennett 1978b, 14-15).

Skinner's experimentele opzet is bedoeld om het intentionele te elimineren, maar maskeert het in feite alleen maar. Hij heeft geen niet-intentionele gedragswetten gevonden. Zijn voorspellingen zijn ook intentioneel (de muis wil kaas en meent kaas te krijgen door op het pedaal te duwen - waarvoor ze goede redenen heeft - en duwt dus op het pedaal), maar dat is versluierd omdat de situatie maar één lichamelijke beweging toelaat.

Skinner heeft gelijk als hij meent dat intentionaliteit geen basis voor psychologie kan zijn, en ook als hij naar puur mechanistische verklaringen zoekt. Maar er is geen reden om zulke verklaringen te zoeken op het persoonlijke niveau, op de oppervlakte van molair gedrag (Dennett 1978b, 14-15). Intensionele zinnen kunnen nu eenmaal niet vertaald worden in extensionele zinnen (Chisholm 1957). Skinner denkt dat het wel kan, maar dat lijkt hem alleen zo omdat hij de intentionele verbanden in een artificiële dwangbuis stopt, of mensen gelijk stelt met wespen (Dennett 1978b, 61-70).

Wat kunnen we doen als we zoeken naar mechanistische, niet intentionele verklaringen, maar als dat onmogelijk is op het persoonlijke niveau? We moeten afdalen naar het sub-persoonlijke niveau, we moeten kijken naar het ontwerp. De cognitiewetenschappers

doen weinig anders. Neem bijvoorbeeld Chomsky's grammatica, misschien wel hèt voorbeeld voor de cognitiewetenschap. Volgens Chomsky is een taalgebruiker in het bezit van een grammatica, dat wil zeggen, hij is in het bezit van een aantal regels met behulp waarvan hij naar believen welgevormde zinnen kan produceren. Het 'bezit' waar hier sprake van is is niet een onschuldige notie van opslag; het gaat hier om 'epistemisch bezit'. Een monolinguale Fransman die de Encyclopedia Brittanica bezit, kan niets doen met al die informatie; maar een taalgebruiker gebruikt zijn grammatica bij iedere zin die hij uit. Het is dus niet genoeg als de regels van de grammatica ergens gerepresenteerd zijn. Ze moeten ook begrepen en uitgevoerd worden.

Nu zijn de regels voor een Chomskyaanse grammatica behoorlijk ingewikkeld, zeker voor complexe zinnen. Het kost mij (MM) nogal wat tijd om de juiste transformaties uit te voeren om bij een complexe zin van een dieptestructuur te komen naar een oppervlakte- structuur. Toch gaat het praten de meeste mensen gemakkelijk af. En dat terwijl het overgrote deel van 's werelds taalgebruikers nog nooit van transformaties, diepte- of oppervlaktestructuren gehoord heeft, en die regels voor transformatie ook niet zou kunnen begrijpen.

Het is volgens de theorie dan ook niet zo dat mensen hun grammatica bewust gebruiken. Weliswaar zijn volgens Chomsky de regels expliciet gerepresenteerd, maar dat betekent niet dat ze bewust zijn. (Waren ze het maar, dan zou het formuleren van die grammatica geen enkel probleem zijn.) 'Bewust' en 'expliciet gerepresenteerd' zijn twee onafhankelijke noties. In feite kun je zeggen dat het niet de taalgebruiker is die de regels van de grammatica gebruikt. Het is zijn taalsysteem dat de transformaties uitvoert, een deel van de taalgebruiker op sub-persoonlijk niveau. De expliciete representaties van de grammatica zijn geen representaties *voor de taalgebruiker,* maar voor een instantie *in* de taalgebruiker.

Zo ook is het niet de luisteraar die een Fourrier-analyse uitvoert op de het oor binnenkomende signalen. Ik zelf zou zo'n analyse niet uit kunnen voeren. Toch worden, volgens de fysiologen, Fourrier-analyses uitgevoerd *in* mij, namelijk door mijn gehoorsysteem.

Voor cognitiewetenschappers is het postuleren van sub-persoonlijke systemen die, voor de persoon onbewust, allerlei computaties en operaties uitvoeren, een natuurlijke zaak. Die sub-persoonlijke

systemen worden meestal in intentionele termen beschreven. Iedere keer dat een theorie-bouwer voorstelt een gebeurtenis, toestand, structuur, enz. in een systeem een *signaal* te noemen, of een *boodschap* of een *instructie* of anderszins er inhouden aan toekent, 'smokkelt' zij nog. Ze poneert impliciet, samen met haar signalen, boodschappen en instructies, iets dat kan dienen als een *signaal-lezer,* een *boodschap-begrijper,* een *instructie-opvolger,* anders dienen haar signalen tot niets, en verdwijnen ze onontvangen en onbegrepen. Deze lezers en begrijpers moeten uiteindelijk gevonden en weggeanalyseerd worden; anders bevat de theorie onder zijn elementen ongeanalyseerde homunculi met genoeg intelligentie om de signalen te lezen, en dan zal de theorie alleen maar het antwoord uitstellen op de grote vraag: wat leidt tot intelligentie? De intentionaliteit van al dit soort gepraat over signalen en instructies herinnert ons eraan dat rationaliteit vooronderstelt wordt, en toont op deze manier aan waar een theorie incompleet is. Zoals Dennett zegt:

> "... wherever a theory relies on a formulation bearing the logical marks of intentionality, there a little man is concealed" (Dennett 1978b, 12).

Dit zou leiden tot een oneindige regressie van homunculi, als het niet mogelijk zou zijn dat er uiteindelijk boodschappen en signalen en instructies zijn die zichzelf begrijpen, en niet meer door weer een andere instantie begrepen of geïnterpreteerd hoeven worden. Volgens Dennett is dit laatste mogelijk geworden sinds de komst van de Artificiële Intelligentie. Daar is immers sprake van representaties die uiteindelijk *tacit* zijn, en zichzelf begrijpen.

We kunnen spreken van zichzelf begrijpende representaties, die een eind aan de regressie van homunculi maken, wanneer we de AI zien als een *top-down* theoretisch onderzoek. Men begint in de AI met een specificatie van een hele persoon of cognitief organisme - neutraal genoemd, een intentioneel systeem - of van een segment van de vaardigheden van die persoon, bijvoorbeeld schaken, of het beantwoorden van vragen over honkbal. Vervolgens wordt dat grootste intentionele systeem opgedeeld in een organisatie van subsystemen, waarvan elk zelf gezien kan worden als een intentioneel systeem, met

zijn eigen gespecialiseerde meningen en verlangens Die subsystemen kunnen dus gezien worden als homunculi.

In feite is er overal in de AI sprake van homunculi, en dat is meestal onschuldig en vaak verhelderend AI-homunculi praten met elkaar, nemen de controle van elkaar over, treden op als vrijwilliger, als klusser, als opzichter, als scheidsrechter Er lijkt vaak geen betere manier te bestaan om te beschrijven wat er gebeurt Een homunculus is alleen gevaarlijk als hij volledig de talenten dupliceert die hij moet verklaren Als men een team of comite kan krijgen van *relatief* onwetende, bekrompen, blinde homunculi om het intelligente gedrag van het geheel te verklaren, dan is er vooruitgang geboekt. Een homunculustheorie is alleen problematisch wanneer in de verklaring van een talent een instantie figureert die datzelfde talent volledig bezit, zoals wanneer de verklaring van hoe we onze schoenveters strikken luidt. "We geven opdracht daartoe aan een schoenstrik-centrum in ons zenuwstelsel"

Een stroomdiagram is typisch het organisatiediagram van een comite van homunculi (onderzoekers, bibliothecarissen, accountants, werknemers); iedere box specificeert een homunculus door een functie voor te schrijven *zonder te zeggen hoe die vervuld kan worden* (men zegt in feite: zet er een mannetje neer om die klus op te knappen). Als we dan de individuele boxen nader bekijken zien we dat de functie van ieder vervuld wordt door hem op te delen via een nieuw stroomdiagram in kleinere, dommere homunculi Uiteindelijk kom je met dit nestelen van boxen in boxen uit bij homunculi die zo stom zijn (ze hoeven alleen maar ja of nee te zeggen) dat ze vervangen kunnen worden door een machine. Men maakt alle homunculi onschadelijk door legers van dit soort idioten te organiseren die het werk doen (Dennett 1978b, 123-124) (53).

Het lijkt er soms op alsof Dennett wil beweren dat met dit opdelen van systemen in subsystemen - van homunculi in subhomunculi - de *intentionaliteit* verdwijnt Dennett gebruikt de begrippen 'intentionaliteit' en 'intelligentie' of 'rationaliteit' ongegeneerd door elkaar. Hij wil de begrijpers van interne representaties weganalyseren, omdat er anders sprake is van homunculi met genoeg *intelligentie* om de representaties te lezen. Daarom wil hij de homunculi steeds dommer maken. Maar met het analyseren van een systeem in steeds dommere

homunculi beantwoordt Dennett wel de vraag: "What makes for intelligence?", maar zeker niet de vraag. "What makes for intentionality?" Immers, ook wat de domste homunculi doen wordt, in een stroomdiagram, nog beschreven in intensionele zinnen, en ze moeten begrijpen waar ze "Ja" of "Nee" tegen zeggen. Dennett doet het soms voorkomen alsof hij intentionaliteit wil verklaren in termen van complexiteit, en intentionaliteit gelijkstelt met intelligentie, verschillende auteurs lezen hem zo (zie b.v. Margolis 1980, Fodor 1981a, Cummins 1983, Van Heerden en Van Zeytveld 1985).

En inderdaad is hier Dennett verre van duidelijk In zijn verhaal over een hierarchie van homunculi lopen twee zaken door elkaar. Ten eerste kun je inderdaad vaak *intelligentie* analyseren in termen van complexiteit. je stelt dan een hierarchie van stroomdiagrammen op om de uitvoering van een intelligentie-vereisende taak op te delen in uitvoeringen van taken die geen intelligentie vereisen En ten tweede kun je laten zien dat de zinnen van de programmeertaal (= bijna natuurlijke taal) via een aantal tussenstappen vertaald kunnen worden in zinnen in de machinetaal, die, in zekere zin, geinterpreteerd is en niet nog eens door de machine 'begrepen' hoeft te worden

Maar we hebben gezien in 4 6.1 dat wat in de machinetaal geinterpreteerd is alleen de instructies voor machinebewegingen (operaties) zijn. Waar die operaties over gaan, de interne representaties op de verschillende adressen, blijft *ongeïnterpreteerd* en onbegrepen Die representaties representeren niets voor het systeem, en zijn dus zeker niet zichzelf begrijpend.

Ofschoon Dennett Fodor's artikel over procedurele semantiek niet gelezen schijnt te hebben, en het in ieder geval nergens noemt (het is in hetzelfde jaar gepubliceerd als zijn boek), weet hij toch wel dat er problemen gezien kunnen worden rond een procedurele semantiek. Hij laat zich op dit punt erg voorzichtig uit:

> "One never quite gets completely self-understanding representations (unless one stands back and views all representation in the system from a global vantage point), but all homunculi are ultimately discharged" (Dennett 1978b, 124)

Als we Dennett lezen alsof hij met zijn hierarchie van homunculi *intentionaliteit* wil verklaren, dan is zijn theorie zeker niet houdbaar Maar we hoeven Dennett niet zo te lezen Voor het probleem van intentionaliteit heeft hij immers zijn notie van een intentioneel systeem ontwikkeld Men kan of tegenover een systeem de fysische houding aannemen, waarbij het gedrag van het hele systeem tot en met dat van het kleinste subsysteem in fysische termen wordt verklaard, of men kan de intentionele houding aannemen, waarbij het gedrag van het hele systeem tot en met dat van het kleinste subsysteem in intensionele termen wordt verklaard Nergens is er ooit een overgang van het niet-intentionele naar het intentionele Maar dat geeft niet in Dennett's theorie want intentionaliteit is *in the eye of the beholder.* Wij zijn het die intentionaliteit aan anderen (intentionele systemen) toeschrijven, om hun gedrag te kunnen verklaren, het systeem heeft die intentionaliteit niet, dus die intentionaliteit hoeft ook niet weggeanalyseerd te worden Het systeem heeft wel intelligentie, en die moet verklaard worden en weggeanalyseerd worden Zo kan men Dennett ook lezen, maar hij blijft op dit punt nogal onduidelijk

Een filosoof kan de zaak volgens hem op twee manieren bekijken men kan toegeven dat de AI inderdaad werkt met zichzelf begrijpende representaties, of men kan de verschillen benadrukken tussen AI representaties en prototypische of *echte* representaties (menselijke uitspraken, kaarten, schilderijen) en concluderen dat de AI niet met echte interne representaties werkt In het eerste geval kun je benadrukken dat de AI-machines bepaalde taken uitvoeren en zich daarbij baseren op computaties in een interne taal Op het niveau van de machinetaal zijn de interne representaties geinterpreteerd omdat ze automatisch leiden tot de uitvoering van de beschreven procedures, ze zijn dus zichzelf-begrijpend en er is geen andere instantie nodig die ze weer interpreteert In het tweede geval kun je benadrukken dat er geen sprake is van echte representaties omdat voor de causale werking van de machinetaal de interpretatie van interne representaties er niet toe doet De machinetaal is in zekere zin geinterpreteerd, maar dat is niet de interpretatie die bedoeld was in de programmeertaal (zie 4 6 1). De lezer die ik plotseling "Schrik!" toebrulde schrok, maar niet omdat hij mijn bevel had geinterpreteerd en begrepen Als men het tweede alternatief kiest, is er nog niets aan de hand de successen

van de AI laten zien dat de psychologie geen interne representaties nodig heeft. Interne pseudo-representaties zijn even goed.

Het probleem van de psychologie is volgens Dennett altijd het volgende geweest: de enige psychologie die gedrag kan verklaren moet interne representaties postuleren Maar interne representaties leiden tot homunculus-regressie (Ryle's argument). Dennett ziet in de AI een kop-ik-win-munt-jij-verliest oplossing Of je ziet AI representaties als zelf begrijpend, en dan is het homunculus-probleem opgelost, òf je ziet ze niet als echte representaties, maar dan blijkt ook dat echte representaties niet nodig zijn, en is het homunculus-probleem ook opgelost.

Wat Dennett voorstelt is een functionalistische psychologie op sub-persoonlijk niveau. Hij is niet bang voor homunculi omdat niet alle interne representaties expliciet hoeven te zijn; op het laagste niveau zijn ze *tacit*. En de vraag of je wel kunt spreken van echte representaties, van echte meningen en verlangens, doet zich niet voor omdat we het hebben over intentionele systemen. Bij een intentioneel systeem doet het er niet toe of het systeem echt meningen en verlangens heeft; het is een intentioneel systeem als je er meningen en verlangens aan kunt toeschrijven.

Daarmee is een fysicalistische psychologie mogelijk geworden. Weliswaar laat een ontwerpbeschrijving, een analyse in de vorm van een stroomdiagram, die door een sub-persoonlijke theorie gegeven moet worden, zich niet uit over de fysische realisatie ervan, maar hij moet wel zo zijn dat een fysische realisatie mogelijk is, hij moet in niet-intensionele taal gesteld kunnen worden. We zullen in de volgende paragraaf Dennett's theorie illustreren aan de hand van een concreet voorbeeld.

5.5. Naar een cognitieve theorie van bewustzijn. Een voorbeeld.

Er gaat heel veel op sub-persoolijk niveau om waarvan de persoon zich niet bewust is, dat ze niet ervaart. Ik heb geen toegang tot de Fourrier-analyses die in mijn gehoorsysteem worden uitgevoerd; ik weet daar niets van, ervaar daar niets van. Uit mijn molair gedrag blijkt dat ze zijn uitgevoerd, maar als die analyses al ergens

gerepresenteerd zijn, dan zijn ze niet voor mij gerepresenteerd. Ik heb geen toegang tot die analyses; delen van mij wel Voor mij heeft de vraag: "Hoe is het om een Fourrier-analyse uit te voeren" geen betekenis.

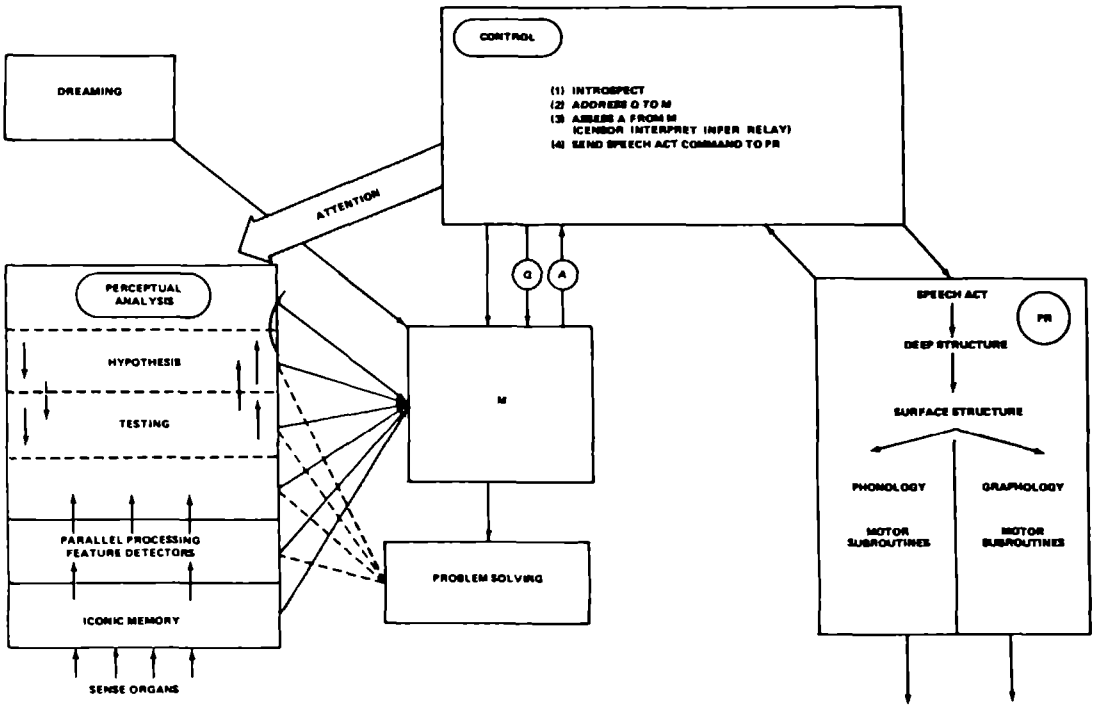
In heel veel gevallen heeft Nagel's vraag: "Hoe is het om een X te zijn?" (naar 'What is it like to be a bat?' Nagel 1974) geen betekenis. Maar in sommige gevallen wel. Er gaat veel in mij om waar ik me niet van bewust ben, maar ook veel waar ik me wel van bewust ben, waarover ik introspectieve en retrospectieve verklaringen kan afleggen Tot die dingen heb ik wel toegang. Dat zijn de kwalitatieve ervaringen die ik noem als men mij vraagt: "Hoe is het om jou te zijn?"

Dennett is, in tegenstelling tot bijvoorbeeld Fodor, niet van mening dat het functionalisme dit soort kwalitatieve kwesties niet aan kan. Hij stelt in een artikel uit 1978 'Toward a cognitive theory of consciousness' (herdruk in Dennett 1978b) voor: "to construct a full-fledged 'I' out of sub-personal parts" (Dennett 1978b, 154). Dat gaat als volgt.

Een persoon is een organisatie van sub-persoonlijke delen. Veel van die sub-persoonlijke delen hebben toegang tot de uitkomsten van elkaars berekeningen, zoals in het voorbeeld van de Fourier-analyse. Zo gaat dat ook in computers: de subroutines hebben toegang tot elkaars output. Men zou kunnen spreken van computationele toegang. Maar voor de persoon zijn dit soort uitkomsten niet toegankelijk Er is geen persoonlijke toegang. Bij een computer heeft de programmeur ook niet automatisch toegang tot de output van subroutines Hij ziet alleen de output van het programma, zeg maar het uiteindelijk gedrag. Toch zou het wel eens handig zijn enigszins te kunnen bijhouden wat de computer aan het doen is. De programmeur kan zijn programma zo schrijven dat er ook informatie wordt afgedrukt over de tussenstappen in de operaties. Er is dan publieke toegang tot die tussenstappen. Men kan voor dit soort publieke toegang zorgen door een afdruk-subroutine te maken die *computationele* toegang heeft tot die uitkomsten waartoe men publieke toegang wil.

Mensen zijn met zo'n computer vergelijkbaar. Er gebeurt van alles in ze waar anderen en zijzelf geen toegang toe hebben. Het publiek ziet alleen het uiterlijk gedrag Maar je kunt iemand wel een aantal dingen vragen waar ze een introspectief of retrospectief antwoord op geeft.

Spreken en schrijven zijn de afdruk-faciliteiten van mensen; en waar we ons bewust van zijn is dat wat we kunnen (na)vertellen. Dennett geeft het volgende stroomdiagram van een 'ik' opgebouwd uit sub-persoonlijke componenten (zie figuur 4).



Figuur 4 (overgenomen uit Dennett 1978b, 155)

Aan het output-einde is een afdruk-component die hij *P(ublic) R(elations)* noemt. PR krijgt als input een opdracht tot het uitvoeren van een taaldaad, of een *semantische intentie;* als output geeft PR een uitspraak. De inrichting en werking van PR is iets wat druk besproken wordt in de psycholinguistiek Laten we even aannemen dat PR gebruik maakt van een Chomskyaanse transformationele generatieve grammatica.

PR krijgt al zijn opdrachten van een zeer machtige component, *Control,* en heeft, via Control, toegang tot een geheugen M.

Stel nu dat Control 'besluit' om 'introspectie' te plegen,

1) hij gaat in de introspectie-subroutine,
2) hij richt een vraag tot M
3) als er een antwoord komt, kan hij
   a) het antwoord censureren
   b) het antwoord interpreteren in het licht van andere informatie
   c) inferenties uit het antwoord afleiden
   d) het antwoord direct naar PR doorsturen
4) de uitkomst van elk van a-d kan een spreekopdracht voor PR vormen.

Het is een noodzakelijke maar niet voldoende voorwaarde voor de inhoud van een opdracht voor PR dat deze als informatie in M zit. Wat voor informatie zit er in M? M krijgt informatie uit de Perceptie-component. De perceptuele analyse vindt plaats op verschillende niveaus, vanaf de stimulatie van de zintuigen tot hogelijk geïnterpreteerde informatie over de waargenomen wereld. Dennett steunt hierbij op de perceptie-theorieen van Neisser (1967). Op het laagste niveau is het iconisch geheugen waar de binnenkomende stimuli vrijwel ongeïnterpreteerd worden opgeslagen. De *feature detectors* geven, in een parallel verlopend proces, informatie over hoeken, randen, lijnen en kleuren. Dan worden hypothesen gevormd en getest, deels gestuurd door kennis-van-de-wereld, van boven, en deels door de data, van onderen, en geholpen door *Problem Solving.* Control regelt de details van al deze processen en verdeelt de aandacht.

De perceptuele analyse-component stuurt informatie naar M vanuit vele niveaus. Immers, wanneer men een complexe scene ziet en die analyseert als, bijvoorbeeld, een tafel met stoelen midden in een

kamer, dan ziet men meer dan alleen dat daar een tafel met stoelen staat Men ziet vormen, kleuren, plaatselijke details en achtergrond.

Perceptie zendt dus informatie naar M Dromen en Probleem-oplossen doen dat ook En ook Control zelf stuurt informatie naar M: zijn doelen, plannen, intenties en meningen. Het is duidelijk dat alle vier componenten, Perceptie, Dromen, Probleemoplossen en Control, niet al hun informatie naar M sturen.

Er gebeurt veel waar PR geen toegang toe heeft, waar de 'ik' niets over kan zeggen. veel intenties en meningen van Control zijn onbewust, evenals vele stappen in de perceptuele analyse of in Probleem-oplossen. De afwezigheid van introspectieve aanwijzingen dat een bepaalde analyse uitgevoerd is is nooit bewijs dat zo'n analyse niet is uitgevoerd De analyse in kwestie kan eenvoudigweg een van de vele processen zijn die op andere wijze bijdragen aan Control, Perceptie en Probleem-oplossen zonder hun resultaten naar M te sturen

Dit was een (heel korte) beschrijving van Dennett's stroomdiagram voor een 'ik' Hij is zich ervan bewust dat zijn control 'awfully fancy' (Dennett 1978b, 164) is, en misschien onmogelijk gebouwd kan worden Maar waar het om gaat is dat het hele systeem een 'ik' moet zijn, terwijl geen van de componenten, ook Control niet, een 'ik' is

Stel nu dat een systeem volgens dit stroomdiagram gebouwd wordt. Hoe zou het zijn om zo'n systeem te zijn? Op het eerste gezicht zou je zeggen. het is helemaal niets om zoiets te zijn, die vraag kun je niet stellen

> "And yet to us on the outside, watching such an entity, engaging it in conversation, listening to its efforts to describe the effects on it of various perceptual environments, there will be at least the illusion that it is like something to be the entity In fact it will tell us (or at least seem to be telling us) just what it is like" (Dennett 1978b, 164-165)

Maar binnenin het machien is niets, geen geest, geen introspecterend oog, alleen duisternis. Binnenin onze schedel is ook duisternis, de processen in onze hersenen gebeuren onwaargenomen en

onwaarnemend.

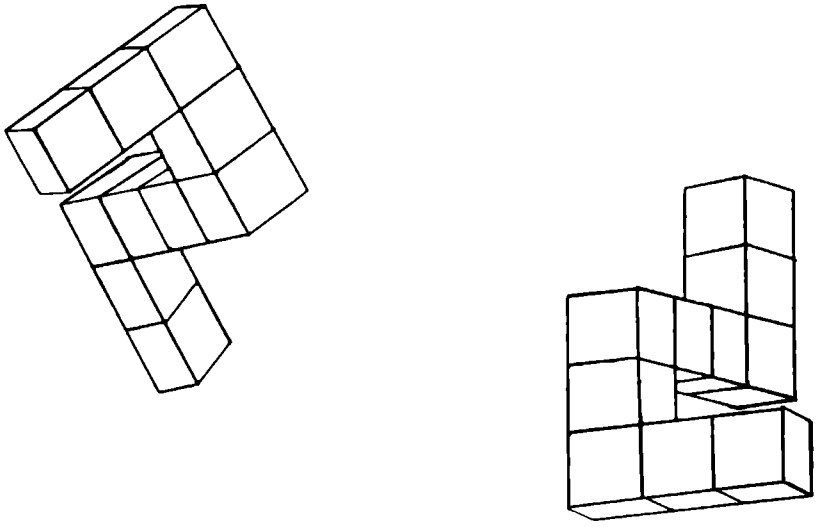
Stel nu, vervolgt Dennett, dat wij een realisatie van dit stroomdiagram zijn. Wat voor persoonlijke toegang hebben wij, en tot wat? Hij noemt een voorbeeld van Lashley wanneer ons gevraagd wordt een gedachte in dactylische hexameter te denken kunnen we (vele van ons, sommigen, steeds minder) dat doen, maar we zijn ons er niet van bewust hoe we dat doen, het *resultaat* schiet ons te binnen, en zover gaat onze persoonlijke toegang. Lashley schijnt gezegd te hebben dat geen activiteit van de geest ooit bewust is, en Dennett interpreteert dit zo dat we bewuste toegang hebben tot de resultaten van onze mentale processen, maar niet tot de processen zelf.

Dennett maakt dan een vergelijking met Hume's analyse van causaliteit. Vroeger meende men dat we een oorzaak en gevolg zagen, en het noodzakelijke verband tussen beide, en daarna en daarom het gevolg afleidden en verwachtten bij het zien van de oorzaak Volgens Hume was dit achterstevoren. Er valt geen oorzaak-gevolg verband te zien; het gaat precies andersom. We zijn geconditioneerd om het gevolg te verwachten, en dat geeft aanleiding tot de illusie dat we een noodzakelijk verband zien dat onze verwachting fundeert en verklaart. De verwachting zelf komt, epistemologisch en psychologisch, eerder.

> "I am proposing a paralell account of "introspection". we find ourselves *wanting to say* all these things about what is going on in us; this gives rise to *theories* we hold about how we come to be able to do this - for instance, the notorious but homespun theory that we "perceive" these goings on with our "inner eye", and that this *perception* grounds and explains the semantic intentions we have" (Dennett 1978b, 166-167, zie ook b v Nisbett and Ross 1980).

Die waarneming *is* er evenwel niet, volgens Dennett Wij nemen bij voorbeeld geen *mental images*, mentale beelden waar. Wij poneren die alleen maar, omdat we merken dat we allerlei dingen willen zeggen over een bepaalde situatie die we niet op dat moment werkelijk zien Bij een taak waarbij we moeten zeggen of twee tekeningen afbeeldingen van

dezelfde vorm zijn (zie figuur 5) kunnen we zeggen hoe een van de twee er na rotatie uitziet, ook bij een rotatie over slechts een paar graden



Figuur 5 (overgenomen uit Dennett 1978b, 167)

Het lijkt alsof we een beeld in onze voorstelling hebben geroteerd, maar dat is een illusie Een illusie van hetzelfde soort als wanneer we een serie sequentieel flitsende lichtjes zien, en het lijkt alsof één licht beweegt Alles wat we weten is wat we willen zeggen, de input voor onze PR De mentale beelden, die we poneren om die oordelen te verklaren, bestaan gewoon niet

Iets soortgelijks gebeurt bij dromen Wij merken soms, 's morgens bij het ontwaken, dat we *iets willen zeggen,* een verward verhaal, een droom We denken dan dat we gedroomd hebben, dat we 's nachts, in slaap, werkelijk dat verhaal ervaren hebben Maar dat is een illusie: de droom-als-ervaring wordt door ons geponeerd om dat wat we willen zeggen te verklaren Maar alles wat er gebeurt is dat er een input voor PR is uit M, de 'herinnering' aan de droom is er, maar de bijbehorende droom-als-ervaring is er nooit geweest.

Uit Dennett's stroomdiagram van een ik' komen een aantal dingen naar voren die redelijk aansluiten bij onze pretheoretische intuities (voor zover die consistent zijn).

1) Men neemt meer waar dan men ervaart (bewust is) De perceptuele analyse levert informatie die niet allemaal naar M gaat, maar wel in het systeem gebruikt wordt
2) Men ervaart alles wat in M komt Vaak vervaagt die ervaring voordat PR er toegang toe krijgt - voordat we hem onder woorden kunnen brengen (*tip-of-the-tongue* verschijnselen!).
3) Men ervaart op ieder moment meer dan men dan wil zeggen PR maakt een keuze uit M, er is in M op ieder moment nog veel meer dat PR had kunnen uitkiezen om te zeggen (*fringe consciousness!*).
4) Men heeft toegang tot de eigen ervaring via de toegang van PR tot M. We kunnen dus eenvoudigweg zeggen wat we ervaren
5) In zekere zin zijn we oncorrigeerbaar waar het onze eigen ervaringen betreft ".. if we say what we mean to say, if we have committed no errors or infelicities of expression, then our actual utterances cannot fail to be expressions of the content of our semantic intentions, cannot fail to do justice to the access we have to our own inner lives" (Dennett 1978b, 171).

Maar natuurlijk is onze toegang tot ons eigen innerlijk leven alleen de toegang tot M, of nog preciezer, de input vanuit M naar PR. Verder is ons innerlijk leven geheel duister en onzichtbaar voor ons

Het hebben van een innerlijk leven is volgens Dennett's theorie een kwestie van het hebben van een bepaald soort functionele organisatie. Een robot met zo'n organisatie *lijkt* misschien geen innerlijk leven te hebben, in zijn metalen huid Maar schijn kan bedriegen. Een werkelijk bewuste entiteit die precies als wij is maar opereert op een tijdschaal tienduizend keer langzamer dan wij lijkt ook geen innerlijk leven te hebben. We zouden zijn dagenlange uitzendingen niet als uitspraken herkennen, laat staan als geestige, vrolijke, droevige of hartgrondige uitspraken.

Als we ons afvragen of iemand bewust is, denken we aan de vraag of in hem een speciaal licht is Maar dat is een vergissing. Wat zo'n vraag in ons eigen geval beantwoordt zijn, volgens Dennett, overwegingen over wat onze huidige capaciteiten en onze activiteiten zijn. Daaraan weten we of we op tijd t bij bewustzijn waren

(conscious). Dezelfde overwegingen moeten ook voor anderen gelden. Men kan over een entiteit die het stroomdiagram realiseert volgens Dennett twee vragen stellen:

1) Lijkt zo'n entiteit (voor ons) een innerlijk, bewust leven te hebben?
2) Heeft zo'n entiteit *in feite* een innerlijk bewust leven?

Op de eerste vraag is een antwoord mogelijk, het gaat hier om kwesties van ontwerp, of die nu belangrijk of triviaal zijn.

Met de tweede vraag is Dennett niet gelukkig. In zijn theorie van intentionele systemen kan zo'n vraag überhaupt niet gesteld worden. Nu hij hem toch stelt vraagt Dennett zich af of er een alternatief is voor "mere doctrinaire verificationism on the one hand, or shoulder-shrugging agnosticism on the other" (Dennett 1978b, 173). Het is volgens hem het probleem van *other minds*, en hij suggereert als oplossing hiervoor een beetje introspectie: wat hij weet van zichzelf, in het licht van de meest belovende psychologische theorieen, aan feiten die hij aanvoert als grond dat hijzelf in feite bewust is, komt goed overeen met cognitivistisch theoretiseren.

Dit voorbeeld illustreert zowel Dennett's eliminatief materialisme (dromen bestaan niet als ervaringen) en zijn instrumentalisme (hij kiest voor een derde-persoonsbenadering waarbij bewustzijn slechts wordt toegeschreven), als zijn strategie om subpersoonlijke theorieën op te stellen voor het verklaren van de (schijnbare) intentionaliteit van de persoon. In het volgende zal ik die kenmerken van zijn theorie achtereenvolgens bekritiseren.

5.6 Eliminatief materialisme en geprivilegieerde toegang.

In deze paragraaf wil ik laten zien dat Dennett in zijn enthousiasme als eliminatief materialist soms wel erg veel elimineert. Zijn verdediging dat de mentale entiteiten die hij elimineert instrumentalistisch gepostuleerde, theoretische entiteiten zijn, is aanvechtbaar.

Volgens zijn eigen zeggen is Dennett ten aanzien van sommige vermeende mentale entiteiten een eliminatief materialist (Dennett 1978b, XX). Dat betekent dat hij het bestaan van sommige (vermeende) mentale entiteiten ontkent. Dit doet hij als alledaagse mentale concepten en predikaten niet in te passen zijn in een (zijn) wetenschappelijke

theorie, of als hun gebruik in de alledaagse taal niet consistent is Dit gebruik van eliminatie levert Dennett hier en daar wat zure opmerkingen op, bijvoorbeeld:

> "If I have understood him correctly, he advances a perspective within which a theorist is to be guided by ordinary usage only where it suits him (i.e. in those domains in which the ordinary mental language is 'well-behaved' in terms of his purposes in isolating discrete and uniform states or attributes of persons), leaving it behind for, or breaking it down into, theoretically manageable 'subpersonal' components" (Coulter 1982, 15).

en

> "'So be it', some philosophers have said, '... If believers won't behave, we won't have any'" (Fodor 1981a, 101).

Mentale entiteiten die zich volgens Dennett niet goed gedragen, en die dus geëlimineerd moeten worden, zijn niet alleen meningen en wensen, maar ook allerlei ervaringen, bijvoorbeeld dromen en mentale voorstellingen Dennett wijdt hier twee artikelen aan: 'Are dreams experiences?' en 'Two approaches to mental images' (beide herdrukt in Dennett 1978b).

Men kan zich afvragen met welk recht Dennett mentale concepten en predikaten kan elimineren of radikaal herdefinieren Maakt hij zich dan niet schuldig aan (Ryleaanse) *category mistakes?* Zo heeft Searle, naar aanleiding van Dennett's intentionele systemen, opgemerkt:

> "The study of mind starts with such facts as that humans have beliefs, while thermostats, telephones, and adding machines don't. If you get a theory that denies this point you have produced a counter-example to the theory and the theory is false" (Searle 1980, 420).

Dennett is niet bang voor dit soort kritiek Onze alledaagse categorieën zijn wel vaker fout gebleken, ze hoeven niet het laatste

woord te hebben. Hij is hierin een medestander van Rorty, de belangrijkste verdediger van het eliminatief materialisme tegen de aantijging van *category mistakes* (b.v. Rorty 1971a, 1971b, 1980), en is het met hem eens als hij zegt naar aanleiding van Searle's opmerking:

> "This brings back memories of the view that the study of the heavens starts with such facts as that the sun moves around in circles and that the earth is at rest" (Rorty 1982, 343).

Dennett geeft toe dat "It would be great fun if Searle and Nagel could be abruptly dismissed as Neanderthal throwbacks..." (Dennett 1982b, 352). Zelf geeft hij vele voorbeelden van alledaagse concepten die verdwenen zijn uit refererend taalgebruik, zoals Satan, heksen, de god Feenoman in een denkbeeldige stam (hoewel *feenomanology* een legitieme vorm van culturele antropologie is, zie Dennett 1978b, 182-186), of die radicaal geherinterpreteerd zijn, zoals atomen, die wel degelijk deelbaar zijn. Wat wij mentale voorstellingen noemen hoeft niet te bestaan, of niet als voorstelling (*non-imagistic*).

Fenomenologen kunnen volgens hem de fenomenen, in dit geval mentale voorstellingen, uitgebreid beschrijven uit de introspectie-protocollen van proefpersonen (54). Zo ook kunnen *feenomanologists* alle eigenschappen van de god Feenoman beschrijven. Maar dat wil niet zeggen dat de fenomenen - of de god Feenoman - echt bestaan. De introspecterende proefpersonen hebben geen directe, geprivilegieerde toegang tot hun innerlijk leven, niemand heeft dat. Dennett heeft dat met zijn stroomdiagram geïllustreerd. We hebben geen mentale voorstellingen, dat is volgens hem een illusie. De mentale voorstellingen zijn alleen in zekere zin gepostuleerd. Wij hebben slechts de input voor onze PR, een serie oordelen, en daaruit leiden we af dat er mentale voorstellingen zijn. Maar dat kunnen we niet weten, we zijn daarin niet oncorrigeerbaar, we hebben geen geprivilegieerde toegang tot zoiets innerlijks als mentale voorstellingen. De mentale entiteiten die Dennett wil elimineren zijn volgens hem ten onrechte gepostuleerde theoretische entiteiten, gepostuleerd op basis van de theorie die inherent is aan ons alledaags taalgebruik.

Tegen deze Rortyaanse verdediging van Dennett, die de geldigheid van het alledaagse taalgebruik aanvalt, is opnieuw kritiek te leveren. Ons alledaagse taalgebruik is geen wetenschappelijke theorie (Bernstein 1971), en het is bijvoorbeeld nooit zo geweest dat een (onware) *theorie* over ondeelbare atomen een rol speelde in de constitutieve *identificatie* van de fenomenen (Coulter 1982). De theorie speelde een rol in de verklaring van de fenomenen, in *hoe* we de fenomenen zagen, maar niet in de identificatie ervan, niet in *dat* we ze zagen. wil elimineren onwaarneembare, gepostuleerde theoretische entiteiten zijn, dan postuleren we ze op grond van een theorie die de (waarneembare) fenomenen moet verklaren. Maar als allerlei ervaringen zelf niet meer als (waarneembare) fenomenen gelden, maar (onwaarneembare) gepostuleerde theoretische entiteiten zijn, hoe kan dat dan? Dennett beweert bijvoorbeeld dat de roterende mentale beelden gepostuleerd zijn door ons, zoals we bewegingen postuleren als we een discrete serie knipperende lichtjes naast elkaar zien. De feitelijke beweging nemen we niet waar, en is door ons gepostuleerd, maar de *ervaring van beweging* is dan toch niet gepostuleerd? We ervaren beweging, en postuleren op grond daarvan dat er beweging is. Die beweging is er niet echt, maar de ervaring is er toch wel echt? Het leek nu juist een van de weinige zekerheden in het leven dat we ons in het bestaan van een aantal ervaringen of gewaarwordingen, zoals van pijn, dromen, mentale voorstellingen, niet konden vergissen, dat die *niet* gepostuleerd en dus feilbaar zijn, maar gegeven. Wij *hebben* onze mentale voorstellingen, wij 'zien' ze, en leiden ze niet af uit een theorie over onszelf.

Rorty ontwikkelt in zijn *Philosophy and the mirror of nature* (1980) de visie dat er *niets* gegeven is, maar kan dan ook niet meer stellen dat de ene theorie beter kan zijn dan de andere. als er niets gegeven is hoeven er ook geen fenomenen of gegevens verklaard te worden. Verschillende theorieën zijn dan alle gelijkwaardige taalspelen (55). Dennett wil evenwel niet zover gaan, volgens hem is er wel iets mentaals gegeven, namelijk de ervaring iets te willen zeggen, de input voor onze PR. We moeten op zijn gezag aannemen dat dat de *enige* ervaring van iets innerlijks is.

Alleen op zijn gezag inderdaad, want als we eens in concreto bekijken op grond waarvan hij beweert dat dromen niet ervaren

worden, dan blijken zijn argumenten niet veel om het lijf te hebben

Om te beginnen oppert Dennett in zijn dromen-artikel alleen de *mogelijkheid* dat dromen geen ervaringen zijn, maar in zijn voorbeeld van een cognitieve theorie van bewustzijn heeft hij die mogelijkheid als actualiteit nodig in zijn stroomdiagram kunnen dromen niet ervaren worden, want er is geen toegang tot dromen, enkel tot de herinnering aan een droom In zijn dromen-artikel kan Dennett niet laten zien dat dromen niet ervaren worden, maar hij wil daar laten zien dat zijn theorie even goed is als de standaard theorie van dromen als ervaringen tijdens de REM-slaap Hij zegt

> It is an *open,* and *theoretical* question whether dreams fall inside or outside the boundary of experience" (Dennett 1978b, 147)

Dennett oppert de theorie dat 's nachts, vermoedelijk gedurende perioden van REM-slaap, een verhaal, een soort droom-cassette, in M wordt geladen, zodat men soms dat verhaal na ontwaken als herinnering kan vertellen Maar dat droomverhaal is nooit 's nachts beleefd Ik vraag mij dan af hoe verklaart Dennett's theorie de fysiologische verschijnselen die zich soms 's nachts voordoen en die passen bij de droominhoud? Hoe verklaart Dennett de snelle oogbewegingen, de nachtmerrie waarvan het angstzweet je uitbreekt en je het soms uitschreeuwt, het praten in de slaap, of de natte droom? Die verschijnselen doen zich ook 's nachts voor als men zich 's morgens geen droomverhaal kan herinneren Hoe kunnen die zich voordoen, als de droom 's nachts niet beleefd is?

Dennett kan deze bezwaren natuurlijk niet over het hoofd gezien hebben Maar kijk eens wat hij ermee doet De snelle oogbewegingen tijdens de REM-slaap die soms correleren met de later herinnerde droominhoud - horizontale bewegingen voor iemand die droomt over een tenniswedstrijd, verticale voor iemand die droomt te oefenen voor basket-ball - maken geen indruk op Dennett. Volgens hem is het even plausibel dat die oogbewegingen optreden tijdens het *onbewuste* laden van een droom in M, als dat ze een min of meer bewuste ervaring van een droom vergezellen Dit lijkt mij al erg onwaarschijnlijk waarom zouden de ogen bewegen op grond van de inhoud van een verhaal dat

een *ontoegankelijk* subsysteem (zie figuur 4) in M aan het laden is?

Bovendien is het schermen met 'bewust' en 'onbewust' in deze context *question begging*. Dennett's belangrijkste argument lijkt te zijn dat tijdens de slaap, dus ook de REM-slaap, de normale fysiologische tekenen van bewustzijn, met name activiteit in de reticulaire formatie van de hersenen, ontbreekt. Maar natuurlijk is dat zo, dromen zijn immers geen *normale* bewuste ervaringen Daar staat tegenover dat tijdens de REM-slaap, in tegenstelling tot tijdens de gewone slaap, het EEG-patroon weer wel gelijkenis vertoont met het waak-patroon met snelle, synchrone, laag-voltage golven (zie Popper and Eccles 1977, Spehlmann 1981). Normaal gesproken zorgt de reticulaire formatie bij het wakker zijn voor de desynchronisering van het EEG-patroon, op grond van externe sensorische stimuli. Tijdens de slaap is het EEG-patroon synchroon. Maar tijdens de REM-slaap is er weer wel desynchronisatie, hetgeen lijkt op het wakker zijn. Anderzijds is de drempel voor *arousal* door auditieve stimuli *hoger* dan tijdens de gewone slaap (visuele externe stimuli zijn er niet omdat de ogen gesloten zijn!). Men noemt de REM-slaap dan ook wel de paradoxale slaap. Dit alles betekent dat Dennett's argument over de reticulaire formatie niet geldt als argument tegen de theorie van dromen als ervaringen. Er is wel een ongevoeligheid voor externe stimuli tijdens de REM-slaap, maar de REM-slaap wijkt zoveel af van de gewone slaap, en lijkt in zoveel opzichten meer op de waaktoestand dan op de gewone slaap, dat niets erop wijst dat een droom *niet* een speciaal soort ervaring zou kunnen zijn.

Maar op de tegenwerping tegen zijn theorie van de emotionele fysiologische verschijnselen die optreden tijdens de REM-slaap reageert Dennett helemaal fantastisch:

> "... the defender of the received view ... can point to the frequent occurrence during REM periods of the normal physiological accompaniments of fear, anxiety, delight, and arousal as considerations in favor of an extended concept of experience How could one exhibit an emotional reaction to something not even experienced? The debate would not stop there, but we need not follow it further now" (Dennett 1978b, 140).

Dennett heeft goede redenen om het debat niet verder te willen volgen: hij wil concluderen dat de strijd tussen zijn theorie en de standaardtheorie *nog onbeslist is*, en vervolgens zijn theorie gebruiken in zijn theorie van bewustzijn. Ik heb ook goede redenen om het debat niet verder te willen volgen: het argument van de emotionele fysiologische verschijnselen tijdens de REM-slaap betekent de definitieve nekslag voor Dennett's theorie van dromen.

Dennett wil allerlei mentale entiteiten elimineren die zich volgens hem niet goed gedragen. Het zal volgens hem niet de eerste keer zijn dat een op het eerste gezicht contra-intuïtieve theorie de verschijnselen pas kan verklaren. De theorie dat de aarde om de zon draait was op het eerste gezicht contra-intuïtief, maar verklaart de verschijnselen, inclusief het feit dat we denken dat de zon om de aarde draait, erg goed. Dennett's theorie van bewustzijn is minstens even contra-intuïtief, maar zijn subtheorie over dromen verklaart in ieder geval de relevante verschijnselen *niet*. Voor zo'n theorie heeft hij het recht niet zoveel te elimineren.

5.7. *Verificationisme.*

Volgens Dennett kunnen we niet echt introspecteren; "We find ourselves *wanting to say* all these things about what is going on in us" (Dennett 1978b, 166). En dat geeft ons aanleiding om theorieën te formuleren over hoe we zulke dingen kunnen zeggen: "Omdat we ze introspecteren". Maar dat is gewoon een illusie, aldus Dennett.

Het is dan heel vreemd te zien dat Dennett nog geen twee bladzijden verder in hetzelfde artikel waarin hij bovenstaande zegt zijn lezers uitnodigt te introspecteren, namelijk wanneer hij het heeft over het roteren van mentale voorstellingen:

> "Now how can my view possibly accommodate such phenomena? Aren't we directly aware of an image rotating in phenomenal space in this instance? No. And that much, I think, you can quickly ascertain to your own satisfaction. For isn't it the case that if you attend to your experience more closely when you say you rotate the image you find it

moves in discrete jumps - it flicks through a series of orientations. You cannot gradually speed up or slow down the rotation, can you? But now "look" again, isn't it really just that these discrete steps are discrete propositional episodes?" (Dennett 1978b, 168; zie ook Pylyshyn (1973b, 1979) die beweert dat mentale voorstellingen propositioneel gerepresenteerd zijn, maar hun bestaan en toegankelijkheid niet ontkent).

Afgezien van het feit dat ik Dennett's intuïties op dit punt niet deel, vraag ik me af wat dit anders is dan introspecteren om te zien dat je niet echt kunt introspecteren! Dennett bedoelt misschien dat wanneer je *probeert* te introspecteren, je merkt dat je niet veel ziet. Toch moet hij daarvoor een beroep doen op een soort *monitoring* van de eigen ervaring, een soort inspectie van de eigen ervaring. Het gaat hier uitdrukkelijk *niet* om de inspectie van het eigen gedrag en de eigen uitingen. Dennett zegt niet "we find ourselves *saying*" maar "we find ourselves *wanting to say*". Volgens Dennett is deze inspectie van de eigen ervaring, van wat we voelen dat we willen zeggen, geen introspectie. Maar het is zeker niet het soort inspectie van het eigen gedrag dat volgens Skinner of Ryle leidt tot een oordeel over wat er *in* ons gebeurt. Volgens Skinner (1964) construeren we een intentie (om onze bril te vinden) door middel van een inspectie van ons openlijk gedrag (rommelen tussen de papieren op het bureau) en een herinnering aan wat dat gedrag bij vorige gelegenheden inhield (bij vorig rommelen zochten we onze bril). Volgens Dennett daarentegen moeten we iets inspecteren dat niet met onze zintuigen waarneembaar is, iets in ons. Het wordt dan wat moeilijk om het verschil met introspectie nog te zien. We moeten wederom uitsluitend op Dennett's gezag aannemen dat dat verschil er is.

Rorty valt Dennett hierop aan (Rorty 1982). Volgens hem durft Dennett niet ver genoeg te gaan. Dennett's hele theorie van intentionele systemen is verificationistisch en vanuit een derde-persoons perspectief gesteld. Bij een intentioneel systeem wordt geen onderscheid gemaakt tussen systemen die echt meningen en verlangens hebben en systemen die ze niet hebben; het zijn beide intentionele systemen als men er meningen en verlangens aan kan toeschrijven op

pragmatische gronden Dennett houdt niet van verschillen die geen verschil maken. Zijn theorie impliceert dat zulke verschillen niet bestaan, en aangezien hij introspectie als bron van informatie afwijst, bestaan er volgens hem geen verschillen die niet voor een derde persoon verschil maken Maar soms schrikt hij terug voor het verificationisme Hij staat sceptisch tegenover Nagel's bewering dat ook vleermuizen een volledig fenomenologisch bewustzijn hebben "without thereby becoming the Village Verificationist" (Dennett 1978b, 152) En wanneer hij de vraag moet beantwoorden of een robot, gebouwd volgens zijn stroomdiagram in feite een innerlijk bewust leven heeft, denkt hij dat er een beter antwoord is dan "mere doctrinaire verificationism" Dan valt hij, net als bij zijn argument over mentale voorstellingen, terug op inspectie van de eigen ervaring, op een eerste-persoons perspectief. Hij wil dan nagaan wat men weet van zijn eigen bewustzijn. Rorty meent dat Dennett deze stap niet kan maken, en denkt dat een verificationisme toch de juiste strategie is. Dennett antwoordt op deze kritiek van Rorty als volgt:

> "I no longer find it polemically useful to insist that I am not any sort of verificationist, with professor Rorty cheering me on (and Putnam offering similar encouragements in recent remarks), I am ready to come out of the closet as some sort of verificationist; but not, please, a Village Verificationist; let's all be *Urbane* Verificationists" (Dennett 1982b, 355).

Ik zou denken dat dit niet iets is om lichtzinnig over te doen. Verificationisme, hoe urbaan ook, lijkt me "a cliff over which to push one's opponent".

5.8. Waar is Dennett?

Dennett besluit zijn boek *Brainstorms* met een artikel, een dessert noemt hij het zelf, getiteld 'Where am I?' Daarin beschrijft hij een science-fiction gedachten-experiment waarbij zijn lichaam gescheiden wordt van zijn centraal zenuwstelsel, dat zijn lichaam radiografisch bestuurt. Bij deze en volgende avonturen heeft hij steeds weer

gelegenheid zich af te vragen: "Waar ben ik? Op de plaats van mijn lichaam of op de plaats van mijn hersenen?" Ik zou ook willen vragen: "Waar is Dennett?", en dan in de zin van: wat is de plaats van Dennett zelf in zijn theorie?

Dennett's theorie over intentionele systemen is een derde-persoonstheorie. Intentionele systemen zijn die systemen waaraan men meningen en wensen instrumentalistisch kan toeschrijven. Hij is geen realist met betrekking tot meningen en wensen. Hij illustreert dit punt aan de hand van de schaakcomputer waaraan de mening kan worden toegeschreven dat hij zijn koningin vroeg in het spel wil brengen. Die toeschrijving geschiedt op grond van het gedrag, door die toeschrijving is het verdere gedrag beter voorspelbaar Maar er is niets in het systeem, geen representatie, die correspondeert met die mening of wens, en er is geen toestand van het systeem die correspondeert met het hebben van die mening of wens.

Een voltooide psychologie mag volgens Dennett echter niet blijven staan bij het verklaren van gedrag in termen van meningen en wensen. Die bestaan niet echt en kunnen dus ook niets veroorzaken. De psychologie dient een ontwerp-houding aan te nemen en het systeem te analyseren in subsystemen - een hierarchie van homunculi. En uiteindelijk moet de ideaal voltooide psychologie een fysische houding aannemen In de intentionele houding schrijven we nog voor het gemak meningen en wensen toe, zoals aan de schaakcomputer (aan wie we trouwens ook het schaakspelen toeschrijven), de computer zelf heeft geen meningen en wensen (en speelt voor zichzelf ook geen schaak). In de fysische houding weten we zoveel dat we die handige fictie niet meer nodig hebben.

Hier wreekt zich de verwarring tussen intentionaliteit-met-een-t en intensionaliteit-met-een-s. Wanneer intentionaliteit gezien wordt als een eigenschap van (derde-persoons) *beschrijvingen* van entiteiten (als intensionaliteit dus), en niet als een eigenschap van die entiteiten zelf, dan kan men inderdaad menen dat intentionaliteit op een gegeven moment kan verdwijnen, als we ophouden te spreken in termen van meningen en verlangens enz. Voor Dennett lijkt dit evenwel ook te betekenen dat we dan ophouden meningen en wensen te hebben. Zoals Margolis zegt:

"The confusion of the issue of psychological realism and of the intensionality of linguistic description offers the only possible explanation of Dennett's remarkably sanguine claim that "the personal story (that is, the 'story' of a person's mental states) has a relatively vulnerable and impermanent place in our conceptual scheme, and could in principle be rendered 'obsolete' if some day we ceased to *treat* anything (any mobile body or system or device) as an Intentional system - by reasoning with it, communicating with it, etc." (1978b, p. 190). Dennett has obviously quite forgotten to eliminate the "we" that do the "treating"" (Margolis 1980, 258; zie ook Leseman 1983, Van Heerden en Van Zeytveld 1985)

Dennett's verhaal is volkomen aannemelijk in het geval van de schaakcomputer, maar men mag toch zeggen dat een computer op het eerste gezicht een dubieus voorbeeld is om de notie van intentionaliteit (in welke zin dan ook) aan te demonstreren (zie ook Fodor 1981a, 103). Maar in Dennett's theorie is een computer juist niet een marginaal geval: dat is net zo goed een intentioneel systeem als mensen Dennett moet, in een uiterst solipsisme (niet te verwarren met Fodor's methodologisch solipsisme) beweren dat alle andere mensen niet anders zijn dan computers, of liever robots. Maar hij kan dat onmogelijk van zichzelf beweren.

Deze kritiek is een variant op een oude kritiek op alle vormen van mechanicisme Dennett noemt deze kritiek heel even aan het eind van een hoofdstuk "Mechanism and responsibility" (Dennett 1978b). Hij vraagt zich dan af of het mogelijk is dat we de intentionele houding helemaal niet meer aannemen ten gunste van de fysische houding Hij bespreekt daarbij een uitspraak van Malcolm, die zegt dat het motto van een mechanicist zou moeten zijn: "Men kan niet spreken, daarom moet men zwijgen". Want om een exclusief mechanicisme te stellen is er tenminste één persoon, namelijk de aangesprokene of in ieder geval de spreker zelf, tegenover wie men niet een exclusief mechanistische houding kan aannemen. Hieruit volgt volgens Dennett niet dat een exclusief mechanicisme niet waar is, enkel dat men het niet kan aannemen, omdat men de intentionele houding niet helemaal kan laten

varen

Dennett maakt hier evenwel een denkfout Het gaat er niet om dat men de intentionele houding niet kan laten varen omdat men die tegenover zichzelf moet innemen, maar het al dan niet aannemen of laten varen van een intentionele of fysische of mechanische houding zelf veronderstelt een intentionaliteit die buiten de theorie van Dennett valt. Waar is in Dennett's theorie degene die een intentionele houding kan innemen tegenover systemen, inclusief het eigen systeem (56)? Dennett zegt zelf:

> "An implication of the view crudely expressed by the slogan that our brains are organic computers is that just like computers their states can be interpreted via a sort of hermeneutical procedure by outside observers to have content - and that's as strong a sort of content as their states can - or could - have We are both the creators and the creatures of such interpretation, and are nothing beyond the reach of that activity" (Dennett 1982b, 355).

Maar hoe kunnen we zelf zowel de scheppers als de schepsels zijn van zo'n interpretatie?

Laten we de zaken nog eens op een rijtje zetten Het probleem van een fysicalistische theorie van het mentale is dat mentale toestanden of gebeurtenissen gekenmerkt worden door kwalitatieve inhoud of door intentionaliteit. Dennett probeert zoveel mogelijk ervaringen als mentale toestanden en gebeurtenissen te elimineren - in 5.6 zagen we hoe. Het probleem van intentionaliteit is dat personen ergens op gericht kunnen zijn, zonder dat datgene waar ze op gericht zijn (zo) in de werkelijkheid bestaat. Het gedrag wordt vaak niet bepaald door fysische stimuli in de fysische buitenwereld, maar door meningen en wensen over die wereld. Personen interpreteren de wereld De standaardoplossing voor dit probleem is het postuleren van interne representaties van de wereld (Fodor's lijn). Dennett lijkt evenwel het probleem gesignaleerd te hebben dat interne representaties door iemand gelezen en geïnterpreteerd moeten worden, en dat dat nooit de persoon zelf kan zijn.

Dennett's oplossing is dat personen geen meningen en wensen

hebben, maar dat wij (of alleen hij?) meningen en wensen instrumentalistisch toeschrijven Zijn suggestie dat in de toekomst, als we meer weten, een fysische houding mogelijk zal zijn om het gedrag te verklaren, duidt erop dat het gedrag volgens hem uiteindelijk dus wel bepaald wordt door de fysische wereld, en niet door een interpretatie van de wereld in de vorm van een interne representatie Maar als er in de fysische wereld geen meningen en wensen bestaan, hoe kunnen wij (of alleen Dennett) dan anderen en onszelf interpreteren als wezens met meningen en wensen? Het probleem van intentionaliteit was dat de wereld geïnterpreteerd wordt, en dat we gericht kunnen zijn op wat niet echt bestaat Dennett lost dit probleem op door te stellen dat intentionaliteit slechts een kwestie van interpretatie is, en niet echt bestaat En daarmee is hij gewoon weer terug bij af

Ook in zijn theorie van bewustzijn zien we dat een volledig intentioneel ik wordt vooropgesteld en niet verklaard Ondanks het elimineren van vele 'interne' ervaringen is er een soort ervaring die Dennett niet kan elimineren de ervaring iets te willen zeggen Maar die ervaring is, zoals Dennett zelf zegt, de input voor PR PR heeft toegang tot die input, of liever, krijgt die input, maar hoe kan het ik daar iets van weten? PR heeft computationele toegang tot de inhoud van M (via Control, Dennett is daar niet duidelijk over, soms zegt hij dat PR toegang heeft tot M, en soms dat alleen Control toegang heeft tot M en daaruit de input voor PR samenstelt), maar hoe kan het ik daar toegang toe hebben? PR *wil* wat zeggen, en het hele systeem, het ik, *zegt* iets Introspectie is immers niet mogelijk, het ik kan niet naar binnen kijken Of misschien is PR zelf het ik. Maar dan hebben we een volwassen en gevaarlijke homunculus op subpersoonlijk niveau

Dennett argumenteert dat vele mentale processen voor ons ontoegankelijk zijn, dat we alleen de resultaten van die processen ervaren Hij doet dat deels heel aannemelijk - wie zou ontkennen dat een aantal interne processen onbewust zijn? Maar die resultaten van interne processen zijn in zijn theorie niet gedragingen, maar nog steeds iets mentaals

Volgens Dennett zijn er geen mentale voorstellingen, en geen dromen-als-ervaring We denken alleen dat ze er zijn Daarmee zijn ze verwezen naar het rijk der illusies, naar wat Rorty, zijn mede-

eliminatief materialist, noemt, "that great dumping ground of out-dated entities, the Mind" (Rorty 1971a, 185) Maar daarmee valt Dennett in een valkuil (57). Hij wil immers een theorie van het mentale geven. Hij kan niet meer zeggen dat mentale voorstellingen of dromen-als-ervaringen 'slechts' mentaal zijn, want dat waren ze al. Dat is een mooie oplossing voor demonen; die kun je verwijzen naar het rijk der illusies, en dan hoef je het bestaan van demonen niet meer te verklaren, want ze bestaan helemaal niet. Maar voor mentale voorstellingen ligt dat anders. Natuurlijk bestaan die niet zoals voorwerpen in de wereld bestaan, ze zijn mentaal. Het maakt niet uit of we ze 'echt' hebben of enkel denken dat we ze hebben - er moet nog steeds verklaard worden waar mentale 'ietsen' zoals illusies vandaan komen Van demonen kun je zeggen: "Het zijn slechts illusies", maar bij mentale entiteiten zelf is dat geen definitieve oplossing.

Dennett spreekt van propositionele episodes, van "thinkings that p". Al bestaan meningen en wensen dan niet, zulke episodische gedachten bestaan kennelijk wel. Ten aanzien van deze expliciete propositionele attitudes kan Dennett niet zeggen dat het "not like anything" is om ze te hebben (zie ook Cummins 1983, die zegt dat in ieder geval voor expliciete propositionele attitudes intentionaliteit nodig is). In zijn theorie is het evenwel onverklaarbaar dat er zulke gedachten zijn. Er bestaat de input voor PR, een opdracht tot een taaldaad. De inhoud van die taaldaad, de propositie die de taaldaad moet uitdrukken, is in die input gerepresenteerd. Gerepresenteerd voor PR, maar niet voor het ik, volgens Dennett's eigen redenering. Zijn stroomdiagram-ik kan wel spreken, maar niet denken, laat staan de herkomst van zijn gedachten interpreteren. Dennett vooronderstelt evenwel ergens een instantie die dat allemaal wel kan, een volledig intentioneel ik, die zijn eigen gedachten als objecten bekijkt en erover gaat nadenken en theoretiseren In zijn stroomdiagram is geen instantie die dat kan; alleen PR kan de gerepresenteerde proposities, de gedachten, 'bekijken', en PR is zelf geen denkend en redenerend subsysteem. Dennett's theorie verklaart geen bewustzijn en geen intentionaliteit, maar vooronderstelt beide

6. CONCLUSIE. INTENTIONALITEIT, FYSICALISME EN DE DUBBEL-ASPECTTHEORIE.

6.1 Inleiding

We hebben in hoofdstuk 1 gezien dat het lichaam-geest probleem zich laat indelen in twee deelproblemen: het qualia-probleem en het probleem van de intentionaliteit Beide problemen vormen een obstakel voor een geheel fysicalistisch wereldbeeld Het qualia-probleem is dat mensen dingen voelen, sensaties hebben, dat mentale toestanden (vaak) een kwalitatieve inhoud hebben Dit probleem vormde de hoofdmoot van de discussie rond de identiteitstheorie. Maar in de cognitiewetenschap speelt dit probleem slechts een marginale rol. (Voor de cognitiewetenschap als cognitieve psychologie is dat niet erg - die gaat over cognitie; maar voor de cognitiewetenschap als filosofie van het mentale is het negeren van dit probleem wel een gemis - zie 3.4.2) In de cognitiewetenschap gaat het om de intentionaliteit van het mentale. Daarbij is het intentionaliteitsbegrip van Brentano overgenomen: psychologische fenomenen worden gekenmerkt door de gerichtheid op een object, maar dat object hoeft niet in de wereld te bestaan, ze zijn gericht op het intentionele object. Met deze notie van de gerichtheid op een intentioneel object kon de cognitiewetenschap recht doen aan een fundamentele intuïtie die door het behaviorisme in de psychologie ontkend werd. namelijk dat het gedrag van een mens (of dier) niet direct bepaald wordt door hoe de wereld is, maar door de manier waarop het organisme de wereld voor zichzelf representeert (zie b.v. Boden 1972, 122, Fodor 1980a, 66). Ondanks deze erkenning van intentionaliteit als kenmerk van het mentale, waarmee dus het mentale wordt onderscheiden van het fysische, wil de cognitiewetenschap toch fysicalistisch zijn. Ze wil een fysicalistische theorie van het mentale geven.

Wanneer men uitgaat van de intentionaliteit van het mentale, heeft men voor een fysicalistische theorie van het mentale twee opties: men kan realist zijn ten aanzien van mentale entiteiten met intentionaliteit, de propositionele attitudes, of men kan instrumentalist zijn ten aanzien van de propositionele attitudes. Men kan natuurlijk bij het opstellen

van een fysicalistische theorie van het mentale helemaal niet uitgaan van intentionaliteit als kenmerk van het mentale, en intentionaliteit niet thematiseren, maar zo'n theorie hoort niet bij de cognitiewetenschap en valt buiten het kader van deze studie (b v. het eliminatief materialisme van de Churchlands of van Rorty (zie noot 55) of het eliminatief behaviorisme van b.v. Watson)

Als voorbeelden van het realisme en het instrumentalisme hebben we Fodor en Dennett bestudeerd. Fodor's realisme ten aanzien van propositionele attitudes samen met een opvatting van intentionaliteit a la Brentano brengen hem ertoe een interne taal te postuleren. Propositionele attitudes zijn relaties tot zinnen in die interne taal, tot interne representaties. Dat leidt tot een probleem waar Brentano ook al mee zat: het probleem van de relatie tussen interne representaties en de wereld. Ik heb dat het referentieprobleem genoemd. Fodor wil ook een fysicalistische verklaring van mentale veroorzaking geven. Hij wil laten zien dat mentale toestanden zowel representationeel zijn als causaal werkzaam. Volgens hem is het 'omdat' in de verklaring 'Hamlet vermoordde de man achter het scherm omdat hij dacht dat het zijn oom was' een *causaal* 'omdat'; hij wil laten zien dat redenen oorzaken zijn. De causale rol van mentale toestanden en processen, van redenen en redeneringen, wordt gespeeld door de interne representaties. Om een causale rol te kunnen spelen moeten interne representaties een fysisch bestaan hebben, ze moeten expliciet in een organisme aanwezig zijn, in de vorm van een of andere neurale structuur. De theorie van het mentale van de cognitiewetenschap is computationeel. Dat betekent dat het de *vorm* van de mentale representaties is die een rol speelt in de causale keten die uitmondt in gedrag (de 'formaliteitsconditie', zie 4.3). En dat leidt tot het tweede probleem.

Als de interne representaties fysische (neurale) structuren zijn, die causaal werkzaam zijn op grond van hun vorm, dan is het niet meer vanzelfsprekend dat die structuren een semantische inhoud hebben, dat ze überhaupt iets representeren. Ik heb dit het betekenisprobleem genoemd. Noch de procedurele semantiek, noch de functionele rol semantiek, noch Fodor's en Dretske's causale theorie van representatie kan op bevredigende wijze verklaren wat een unieke interpretatie van de interne representaties vastlegt. Als Fodor's causale theorie houdbaar was geweest, waren het referentieprobleem en het

betekenisprobleem tesamen opgelost. Maar zelfs dan zou het probleem blijven bestaan dat de representaties wel een unieke interpretatie zouden hebben, maar dat daarmee nog niet gezegd is voor wie ze iets representeren. Ik heb dit het intentionaliteitsprobleem genoemd.

Fodor maakt geen onderscheid tussen de drie verschillende problemen, en probeert ze in een klap op te lossen Dit lukt hem niet; zijn theorie blijft intentionaliteit en een interpreterende persoon vooronderstellen.

Dennett heeft (althans een deel van) de problemen gesignaleerd. Hij weet dat de interpretatie van expliciet aanwezige interne representaties problemen oplevert Dus hij ontkent dat er expliciet aanwezige interne representaties zijn. Hij is zelfs een instrumentalist voor wat betreft propositionele attitudes. Zeker, aan een intentioneel systeem schrijven we meningen en verlangens toe. Maar een intentioneel systeem is alles waartegenover we de intentionele houding aannemen. In feite werken alle systemen gewoon volgens fysisch-causale wetten. Verklaringen in termen van meningen en verlangens zijn alleen gemakkelijk om voorspellingen te kunnen doen zolang we nog niet alle fysisch-causale wetten weten of kunnen hanteren. Maar we hebben gezien dat Dennett niet kan verklaren hoe het kan dat tenminste hijzelf een intentionele houding kan aannemen. Hij kan zeggen dat de intentionele houding van het mechanisme Dennett de beste manier is om te reageren op zijn omgeving, maar dat verklaart op zich niet hoe het mechanisme Dennett in staat is anderen en zichzelf als intentionele systemen te beschouwen, en hun en zijn eigen toestanden als toestanden met een bepaalde inhoud te interpreteren. Ook zijn theorie blijft intentionaliteit en een interpreterende persoon vooronderstellen.

Realisme en instrumentalisme voor wat betreft intentionaliteit, propositionele attitudes en interne representaties, zijn de twee enige mogelijkheden die er zijn. Vormen Fodor en Dennett ook de enige twee mogelijkheden? Niet helemaal. Er zijn binnen het fysicalisme nog een paar opties open die ik nog niet heb besproken. (Fodor geeft in zijn artikel 'Fodor's guide to mental representations' uit 1985 een fraaie indeling van de mogelijkheden)

Aan de kant van de realisten is er, behalve de aanhangers van een fusie-theorie van propositionele attitudes (zie 4.2) en van de procedurele of de functionele-rol-semantiek (zie 4.6.1 en 4.6.2) ook

nog Searle Searle is het met Fodor eens dat propositionele attitudes bestaan, en dat intentionaliteit bestaat, maar hij is het verder op de meeste punten met Fodor (en met de hele cognitiewetenschap) oneens Zijn theorie lijdt niet aan de problemen die Fodor heeft Volgens Searle is het object van een propositionele attitude *niet* een intentioneel object, maar een object in de wereld Daarmee ontloopt hij het solipsisme-probleem Hij hoeft ook niet te stellen dat de interne representaties door hun vorm causaal werkzaam zijn, want hij stelt dat een mentale toestand een handeling kan veroorzaken in de zin van intentionele veroorzaking' Voorts beweert Searle dat hij een fysicalist is, en dat intentionaliteit een product is van de *'causal powers'* van de hersenen (Searle 1980, 1983) Ik heb Searle niet uitgebreid behandeld omdat hij weliswaar niet ten prooi is gevallen aan de problemen waar Fodor onder lijdt, maar die problemen op een goedkope manier ontloopt Searle laat namelijk onvoorstelbaar en mysterieus wat Fodor juist probeert uit te leggen Searle stelt gewoon dat mentale veroorzaking plaatsvindt, en dat betekenis causaal werkzaam is, maar hij laat niet zien hoe Als fysicalist zou hij niet een primitieve notie van 'intentionele oorzaak mogen hebben En terecht is hij na zijn artikel 'Minds, brains and programs' (1980) in het tijdschrift *Behavioral and Brain Sciences* aangevallen op zijn volstrekt duistere notie van de *causal powers* van de hersenen, die intentionaliteit zouden afscheiden

Voor de fysicalist die geen realist is voor wat betreft propositionele attitudes en intentionaliteit is er nog een aantal opties open Een wat extreme positie neemt bijvoorbeeld de filosoof Stich in Hij is geen realist maar ook geen instrumentalist zoals Dennett Hij is het vrijwel volledig eens met Fodor s theorie, behalve dat hij het geen probleem vindt dat de inhoud van mentale representaties niet door de theorie verklaard wordt, en geen rol speelt in de computaties Hij stelt voor om Fodor s theorie van het mentale met de formaliteitsconditie te handhaven en de hele notie van inhoud gewoon te laten vallen De mentale representaties zijn dan causaal werkzaam vanwege hun vorm, maar ze representeren helemaal niets (zie Stich 1983) Ik moet zeggen dat ik deze positie niet goed begrijp Door de notie van inhoud te laten vallen vermijdt Stich het probleem van de interpretatie van interne representaties in de filosofie van het mentale Maar hij

verschuift het probleem van inhoud alleen maar naar de taal. Stich kan moeilijk volhouden dat natuurlijke taal ook geen inhoud heeft - dat zou betekenen dat hij helemaal niets kan zeggen. En zelfs als hij het niet met Fodor eens is dat de semantische eigenschappen van de taal afkomstig zijn van die van de interne representaties, dan is dat nog geen reden om te stellen dat interne representaties helemaal geen semantische eigenschappen hebben. Dat zou immers betekenen dat we niet (bewust) kunnen nadenken over wat we willen zeggen of schrijven, of een zin stilzwijgend voor onszelf formuleren. Het feit dat niet verklaard kan worden in de fysicalistische theorie wat de semantische eigenschappen van interne representaties vastlegt lijkt mij een vreemde reden om te concluderen dat de representaties geen semantische eigenschappen *hebben*. Het is al erg genoeg om de baby met het badwater in de afvoer te kieperen, maar Stich gooit het bad en de afvoer zelf ook nog weg (zie ook Fodor's artikel 'Narrow content')

Mijn conclusie is dat er in de cognitiewetenschap geen houdbare fysicalistische theorie van het mentale is; al zulke theorieën moeten intentionaliteit en een interpreterende persoon vooronderstellen. Waarom is dat erg? Dat is erg omdat, ofschoon de besproken fysicalisten in eerste instantie uitgaan van intentionaliteit als kenmerk van het mentale, ze toch vinden dat er iets intrinsiek verkeerd is aan intentionaliteit. Intentionaliteit is geen fysicalistische entiteit of eigenschap of relatie, en mag dus niet een rol spelen in een fysicalistische theorie van het mentale. Daarom zoekt Fodor naar een oplossing voor het probleem van de semantische interpretatie van interne representaties waarvan ". . the vocabulary contains neither intentional nor semantical expressions" (Fodor 1984a, 232). En daarom beweert Dennett: "Intentionality ... serves as a reliable means of detecting exactly where a theory is *in the red*" (Dennett 1978b, 12) (58). Intentionaliteit hoort niet thuis in een fysicalistische theorie. Maar Fodor noch Dennett slagen erin om die vervelende intentionaliteit kwijt te raken.

De filosoof Putnam, wiens artikel 'Minds and machines' in 1960 aan de wieg van de cognitiewetenschap stond, en die een groot voorvechter van het reductionisme is geweest, heeft in zijn *Reason, truth and history* uit 1981 gezegd:

"... 'good', 'right' (and also 'justified belief', 'refers' and 'true') are not identical with physicalistic properties and relations What *this* shows is not that goodness, rightness, epistemic justification, reference and truth do not exist, but that monistic naturalism (or 'physicalism') is an inadequate philosophy" (Putnam 1981, 211).

Wanneer we ook 'intentionaliteit' in het rijtje 'goedheid, juistheid, epistemische rechtvaardiging, referentie en waarheid' plaatsen ben ik het van harte met hem eens. Wat immers zijn de argumenten voor een fysicalisme dat ons lijkt te dwingen om te zeggen dat goedheid enz. en intentionaliteit niet bestaan?

Het fysicalisme is een filosofische vooronderstelling en niet een empirische theorie. De argumenten voor het fysicalisme zijn filosofische argumenten. Volgens velen echter dwingen de wetenschappen ons tot een positie van fysicalisme, en zijn er zelfs empirische aanwijzingen voor het fysicalisme. In het volgende wil ik het gewicht van die empirische aanwijzingen kritisch evalueren

6.2. Directe empirische aanwijzingen voor het token-fysicalisme.

We hebben in 3.3.1 gezien, aan de hand van Fodor's artikel 'Special sciences', dat er zwaarwegende argumenten tegen het type-fysicalisme en vóór het token-fysicalisme zijn. Ik onderschrijf de argumenten tegen het type-fysicalisme, maar wat zijn de argumenten vóór het token-fysicalisme?

Fodor laat zien dat volgens het token-fysicalisme er brugprincipes zijn die een bepaalde psychologische eigenschap identiek stellen aan een disjunctie van neurofysiologische eigenschappen. Hij meent dat er voor het token-fysicalisme empirische aanwijzingen aangevoerd kunnen worden en verwerpt het argument dat de enige mogelijke empirische aanwijzingen voor het token-fysicalisme ook aanwijzingen voor het type-fysicalisme zouden zijn, omdat alleen de ontdekking van psychofysische correlaties tussen typen van gebeurtenissen mogelijk zou zijn (zie ook Davidson 1970). In deze paragraaf wil ik laten zien dat Fodor niet waar kan maken dat er empirische aanwijzingen voor het

token-fysicalisme kunnen zijn.

Eerst Fodor's redenering: er moet aangetoond worden dat de neurofysiologische tegenhangers van type-identieke psychologische gebeurtenissen identiek zijn voor wat betreft die eigenschappen die bepalen wat voor soort *psychologische* gebeurtenis een bepaalde gebeurtenis is. Immers, gebeurtenissen op zich kunnen allerlei soorten eigenschappen hebben, sommige bepalen bijvoorbeeld wat voor *neurofysiologische* gebeurtenis zo'n gebeurtenis is, andere wat voor *psychologische* gebeurtenis. De neurofysiologische tegenhangers van type-identieke psychologische gebeurtenissen zijn verschillend voor wat betreft hun neurofysiologische eigenschappen, niet voor wat betreft hun psychologische eigenschappen. Fodor stelt zich de vraag:

> "*Could* we have evidence that an otherwise heterogeneous set of neurological events have those kinds of properties in common?" (Fodor 1981a, 137).

Zijn antwoord op die vraag is vol vertrouwen: natuurlijk wel! De neurologische theorie zelf kan verklaren waarom een n-tal van neurologisch verschillende gebeurtenissen identiek zijn voor wat betreft hun psychologische eigenschappen, of anders kan een nog fundamentelere wetenschap dat.

Maar dat lijkt nu juist onmogelijk! Die psychologische eigenschappen komen immers juist niet voor in het vocabulaire van de neurofysiologie of meer fundamentele wetenschappen. Hoe kan een *neurofysiologische* theorie verklaren dat verschillende neurologische gebeurtenissen de psychologische sensatie 'rood' gemeen hebben? Hoe kan een *fysische* theorie verklaren dat verschillende fysische gebeurtenissen de eigenschap 'verkoop van een huis' hebben? Het hele argument van Fodor tegen het reductionisme was juist dat dit soort verklaringen in de meer fundamentele wetenschappen niet mogelijk waren, omdat de eigenschappen die voorkomen in de wetten van de 'hogere' wetenschappen juist niet voorkomen in de wetten van de meer fundamentele wetenschappen Voor zover dat wel het geval is, is er sprake van reductionisme.

Bovendien vooronderstelt Fodor de mogelijkheid om een gebeurtenis te identificeren en te individualiseren onafhankelijk van enige

eigenschap. Het valt evenwel in de praktijk nauwelijks aan te geven *welke* gebeurtenis in de hersenen, met bepaalde neurofysiologische eigenschappen, een bepaalde, op zijn psychologische eigenschappen geïdentificeerde gebeurtenis is (zie Davidson 1969).

Putnam (1981) geeft een argument dat laat zien dat niet alleen de identiteit tussen verschillende neurofysiologische gebeurtenissen en een bepaald soort psychologische gebeurtenissen niet empirisch aantoonbaar is, maar dat ook correlatie tussen die twee problematisch is. Dat argument gaat als volgt.

De hersenen vormen een zeer complex systeem van talloze neuronen. Daartussen zijn allerlei causale verbanden. Ieder gebeurtenis in de hersenen heeft velerlei gevolgen in velerlei delen van de hersenen. Er is zelden of nooit sprake van een lineaire causale keten; er zijn vertakkingen en samenlopen, een causaal netwerk. Het probleem is dat de psychologie mentale gebeurtenissen vaak opdeelt op een vrij *discrete* wijze. Hier is een sensatie van rood. Nu is hij begonnen, nu opgehouden. Causale netwerken zoals de hersenen zijn niet discreet. Er is geen unieke fysische gebeurtenis die *het* correlaat is van een sensatie (Putnam 1981, 86; zie ook Coulter 1982) (59). Welke hersentoestanden moeten dan in de disjunctie opgenomen worden? Aan welke disjunctie van hersentoestanden is het hebben van een rood-ervaring identiek? Er zijn een aantal observationeel ononderscheidbare mogelijkheden. Als het token-fysicalisme waar is, valt niet uit te maken op wat voor manier het waar is; valt niet uit te maken aan *welke* hersentoestanden een gegeven sensatie-toestand identiek is. Tijdens mijn rood-ervaring heb ik talloze hersentoestanden. En bij een volgende rood-ervaring weer talloze andere. Welke horen in de disjunctie thuis (60)? Om een correlatie te vinden tussen rood-ervaringen en neurofysiologische gebeurtenissen is op zijn minst een groot aantal (n) bestudeerde rood-ervaringen nodig. Iedere rood-ervaring gaat zonder meer samen met een neurofysiologische gebeurtenis, al was het alleen maar omdat beide soorten gebeurtenis op zijn minst werkende hersenen vooronderstellen. Maar wat te zeggen als, zoals Fodor (1981a, 137) mogelijk acht, ieder van die n rood-ervaringen samen gaat met een ander geheel van neurofysiologische gebeurtenissen? Hoe kun je dan nog spreken van correlatie? Of moet men zeggen dat een rood-ervaring identiek is aan de disjunctie van

alle mogelijke hersentoestanden? Maar wat is dan nog verschil tussen rood-ervaringen en blauw-ervaringen (61)?

Of neem een ander voorbeeld. Volgens Thomas Nagel (1974) is het niet voor te stellen wat het is om een vleermuis te zijn. Maar zou het niet mogelijk zijn dat een vleermuis een rood-ervaring heeft? (N.B Ik heb me laten vertellen dat vleermuizen, in tegenstelling tot de gangbare opvattingen, wel degelijk goed kunnen zien.) Vleermuizen hebben een wat ander neurologisch systeem dan mensen. De volledig gespecificeerde hersentoestanden van vleermuizen wanneer ze naar rode voorwerpen kijken kunnen dus niet identiek zijn aan die hersentoestanden die identiek zijn aan menselijke rood-ervaringen. Stel een vleermuis-rood-ervaring-hebben is identiek aan de disjunctieve eigenschap van de vleermuishersenen 'P1 of P2'. En stel mijn rood-ervaring-hebben is identiek aan de disjunctieve eigenschap van mijn hersenen 'P3 of P4'. Nu zijn er twee mogelijkheden:

1) de kwalitatieve aard van de vleermuis-rood-ervaring is identiek aan de disjunctieve eigenschap 'P1 of P2' en het kwalitatieve karakter van mijn rood-ervaring is identiek aan de daarvan verschillende disjunctieve eigenschap 'P3 of P4'.

2) de kwalitatieve aard van de vleermuis-rood-ervaring is identiek aan de kwalitatieve aard van mijn rood-ervaring en beide zijn identiek aan de complexe disjunctieve eigenschap 'P1 of P2 of P3 of P4' (zie Putnam 1981, 93). Die beide mogelijkheden zijn observationeel niet te onderscheiden.

Al deze problemen geven aan dat betwijfeld kan worden dat voor het token-fysicalisme *empirische* bewijzen aangevoerd kunnen worden. Het is niet duidelijk wat voor correlaties er kunnen bestaan wanneer de rechterkant van de brugformules disjunctief is. Het is al evenmin duidelijk wat voor empirische bewijzen er zijn om bepaalde eigenschappen al dan niet in die disjunctie op te nemen, met andere woorden, het is niet duidelijk wat die disjunctie moet zijn. Zodra we weten dat een bepaald soort psychologische gebeurtenis *steeds* samengaat met een bepaald soort neurofysiologische gebeurtenis is dat een aanwijzing voor het type-fysicalisme. Als die correlatie zo niet bestaat is het onmogelijk te weten met *welk* van de talloze neurofysiologische gebeurtenissen op een bepaald moment een bepaalde psychologische gebeurtenis samengaat. En het is zeker onmogelijk dat

een neurofysiologische theorie kan verklaren dat de verschillende neurofysiologische gebeurtenissen die samengaan met type-identieke psychologische gebeurtenissen enige psychologische eigenschappen gemeen hebben.

6.3. Indirecte empirische aanwijzingen voor het token-fysicalisme.

Het ziet er, gezien bovenstaande argumenten, dus niet naar uit dat er directe empirische aanwijzingen voor het token-fysicalisme kunnen bestaan. Maar misschien bestaan er indirecte empirische aanwijzingen. Een bekende verdediging van het fysicalisme gaat als volgt: Kijk eens naar het succes van het bijbehorende researchprogramma, de cognitiewetenschap. "It's the only game in town!" Het is volstrekt onduidelijk wat een alternatief programma zou kunnen zijn. Zonder fysicalisme is er geen wetenschap mogelijk!

Ook dit argument klinkt krachtiger dan het in feite is. We zagen in hoofdstuk 2 dat er op een belangrijk punt in de cognitiewetenschap *geen* empirische vooruitgang is. Op het gebied van de 'hogere' cognitieve processen, de globale denkprocessen, zijn er grote problemen. Er zijn geen computers die net als mensen zijn, omdat onze globale denkprocessen niet formaliseerbaar en dus niet programmeerbaar blijken. Wanneer men aanneemt dat die problemen van praktische, en niet van principiële aard zijn, dan is dat omdat men ervan overtuigd is dat het fysicalisme waar is, dat de mens een volledig fysisch systeem is, en als zodanig nagebouwd moet kunnen worden. Maar het fysicalisme moet dan op andere, onafhankelijke argumenten steunen, en mag niet zelf weer zitten hopen op - alsmaar uitblijvende - empirische successen. Het is wel een erg immuniserende manoeuvre om problemen bij de empirische vooruitgang als onbelangrijk te zien met een beroep op de onfeilbaarheid van de filosofische kern, en tegelijkertijd problemen in de filosofische kern af te doen als onbelangrijk met een beroep op te verwachten grote empirische successen.

Er is evenwel nog een sterkere tegenwerping tegen het beroep van de fysicalisten op het empirisch succes van de cognitiewetenschap. Dat succes is zeker op een aantal punten wel aanwezig - en misschien is

het wel "the only game in town" of "the only straw afloat" - maar dat succes is geen succes van de cognitiewetenschap *als fysicalistische theorie van het mentale.* Weliswaar beweren de cognitieve psychologen en de cognitiewetenschappers dat ze bezig zijn met een fysicalistische theorie - men beschouwt het fysicalisme als vanzelfsprekend - maar dat is niet meer dan lippendienst. In feite is de cognitiewetenschap door en door intentioneel, dat wil zeggen dat ze voortdurend intentionaliteit vooronderstelt en voortdurend intentionele termen in intensionele zinnen gebruikt om de gebeurtenissen in haar domein beschrijven

Dat is wat Fodor en Dennett juist proberen aan te tonen De cognitiewetenschap analyseert de cognitieve processen in termen van informatie-stroom, van het testen van hypothesen, het trekken van conclusies en het nemen van beslissingen Ook de taken van de domste homunculi worden in intentionele termen beschreven.

Wat Dennett voorstelt is een *hervorming* van de huidige cognitieve psychologie, zodat uiteindelijk het intentionele taalgebruik en de toeschrijving van intentionaliteit eruit zal verdwijnen. Hij zegt·

> "Intentional theory is vacuous as psychology because it presupposes and does not explain rationality and intelligence" (Dennett 1978b, 15)

Deze opmerking is evenwel niet juist. Intentionele cognitieve psychologie kan - met een hierarchie van steeds dommere homunculi - juist wel (het formaliseerbare deel van) intelligentie goed verklaren. Wat vooronderstelt blijft is niet intelligentie maar intentionaliteit. Fodor zegt hierover

> "The point is that machine operations - *including elementary machine operations* - are themselves characterized in ways that involve intensional idiom insofar as their specification is relevant to their role in psychological explanations. For intensionality - as opposed to intelligence - it's (as you might say) a dual aspect theory all the way down, with intensional characterization specifying one of the aspects and mechanical characterization specifying the other" (Fodor 1981a, 22).

Ook Fodor beweert dat de cognitieve psychologie door en door intentioneel is, en dat niemand nog een theorie van intentionaliteit heeft, al vindt hij zijn pogingen om het probleem van de semantische interpretatie van interne representaties op te lossen een stap in de goede richting (b.v Fodor 1985, 99).

De huidige cognitiewetenschap vooronderstelt intentionaliteit en is daarmee geen fysicalistische theorie van het mentale; ze is evenmin gebaseerd op een fysicalistische theorie van het mentale. Het empirisch succes van de cognitiewetenschap kan dus ook geen argument zijn voor het fysicalisme

6.4. De dubbelaspecttheorie en de persoon.

Fodor en Dennett willen de huidige cognitiewetenschap respectievelijk aanvullen of hervormen, omdat ze menen dat 'intentionaliteit' geen basisbegrip, geen primitieve notie, in de psychologie mag zijn. En ze vinden dat dat niet mag, omdat 'intentionaliteit' geen fysicalistische notie is Maar wat zou er gebeuren als dat wel zou mogen?

We hebben gezien in 3.3.1 dat, volgens de kritiek op het type-fysicalisme, de *eigenschappen* die genoemd worden in de speciale wetenschappen geen fysi(calisti)sche eigenschappen hoeven te zijn. De speciale wetenschappen hebben ieder hun eigen indelingen van de dingen of gebeurtenissen in de wereld. Het zou dus niet als een verrassing mogen komen en geen probleem mogen zijn dat intentionaliteit als eigenschap in de psychologie geen fysicalistische eigenschap is.

Het token-fysicalisme, dat zelf de voornaamste kritiek op het type-fysicalisme geleverd heeft, stelt daarnaast dat alle dingen en gebeurtenissen in de wereld *fysische* gebeurtenissen en dingen zijn. Maar op grond van wat kan men dat nog zeggen? We hebben gezien dat er geen directe empirische aanwijzingen voor het token-fysicalisme zijn Het enige wat men kan zeggen is dat 1) alle dingen en gebeurtenissen fysische eigenschappen hebben, 2) sommige dingen en gebeurtenissen mentale eigenschappen hebben en 3) voor mentale eigenschappen geen fysische verklaring gegeven kan worden (zie ook Davidson's *anomalous monism,* 1970). Punt 3 houdt in dat er geen voldoende en noodzakelijke

fysische voorwaarden gegeven kunnen worden wanneer een ding of gebeurtenis een mentale eigenschap heeft (geen brugprincipes, zie 3.3.1) - anders was die mentale eigenschap wel fysisch verklaarbaar. Punten 1 t/m 3 houden tesamen in dat sommige dingen en gebeurtenissen alleen fysische eigenschappen hebben, en dat andere dingen en gebeurtenissen zowel fysische als mentale eigenschappen hebben, terwijl op grond van de fysische eigenschappen niet uit te maken valt welke dingen en gebeurtenissen tot de ene en welke tot de andere categorie behoren, en geen enkele combinatie van fysische eigenschappen een voldoende en noodzakelijke voorwaarde vormt voor het hebben van een mentale eigenschap. Is het in dat geval niet op zijn minst misleidend, of zelfs onjuist, om nog van fysicalisme of van monisme te spreken? Er zijn duidelijk twee soorten dingen of gebeurtenissen, en niet één soort. Weliswaar hebben beide soorten fysische eigenschappen, maar één soort heeft daarnaast ook nog mentale eigenschappen. Als die laatste soort dingen en gebeurtenissen al *fysische* dingen en gebeurtenissen genoemd mogen worden, zijn het toch zeker een heel ander soort fysische dingen en gebeurtenissen dan de eerste soort. Ik zou het duidelijker vinden om de dingen die naast fysische eigenschappen ook mentale eigenschappen hebben geen fysische dingen te noemen, ter onderscheiding van de dingen die alleen fysische eigenschappen hebben.

Intentionaliteit is zo'n mentale eigenschap die niet in fysische termen verklaarbaar is. Als mijn analyse van Fodor's probleem met intentionaliteit en mijn conclusie in 4.7 juist zijn, dan is intentionaliteit niet een eigenschap van mentale representaties, maar van personen. Ook bij Dennett's theorie is altijd nog een persoon voorondersteld die de eigenschap intentionaliteit heeft, hetzij als homunculus, hetzij als de persoon die de intentionele houding kan innemen (zie 5.8). De persoon is het niet-fysische ding dat naast allerlei fysische eigenschappen ook een mentale eigenschap heeft.

Mijn voorstel is om het token-fysicalisme (of het *anomalous monism*) te vervangen door een dubbelaspect- of persoonstheorie à la Strawson (zie 1.2). Volgens deze theorie bestaan er twee soorten entiteiten in de wereld: entiteiten met alleen fysische eigenschappen (fysische dingen), en entiteiten met zowel fysische als mentale eigenschappen (personen). Zoals gezegd, dit is geen Cartesiaans dualisme. De

persoon is niet een compositum van fysisch lichaam en niet-stoffelijke geest, maar eén ondeelbaar geheel, op zichzelf noch fysisch noch psychisch, maar met zowel fysische als mentale eigenschappen. De notie van 'persoon' is in die zin een primitieve notie, een basisbegrip. De persoon heeft een aantal fysische eigenschappen. gewicht, chemische samenstelling, fysiologische werking enz; de persoon heeft ook een aantal mentale eigenschappen Intentionaliteit is zo'n mentale eigenschap, of in een wat andere zin, intentionaliteit is het kenmerk van (althans een deel van (zie 1.4.2)) de andere mentale eigenschappen. Het hebben van meningen en wensen en sensaties enz. zijn mentale eigenschappen van de persoon. Men kan niet vragen "Wat is het om een mening of een wens of een pijn *te zijn*", want er bestaan geen mentale entiteiten als meningen en wensen en pijnen. Die woorden hebben geen referent in de zin van een ding in het lichaam van de persoon. In zoverre heeft Dennett gelijk met zijn eliminaties. Maar het is wel mogelijk om te vragen: "Wat is het om een mening of een wens of pijn *te hebben*", want dat zijn allemaal mentale eigenschappen van de persoon. Meningen en wensen en pijn bestaan wel, maar niet als dingen, en al helemaal niet als *fysische* dingen. Het zijn *eigenschappen*, en deze mentale eigenschappen zijn niet verklaarbaar in termen van of reduceerbaar tot fysische eigenschappen. In sommige gevallen, met name in het geval van sensaties, is het wellicht mogelijk om aan te geven welke fysische eigenschappen *noodzakelijk* zijn voor het hebben van die sensaties, maar zelfs hier zijn er nog problemen (zie Dennett's 'Why you can't make a computer that feels pain', in 1978b).

De psychologie kan de vraag 'Wat is het om een mening te hebben?' beantwoorden met een intentionele theorie over interne representaties, maar ook die representaties moeten niet gezocht worden als fysische dingen in het hoofd van de persoon. Ze hebben misschien wel psychologische realiteit, maar geen fysische realiteit.

6.5. Mogelijke bezwaren tegen een dubbelaspecttheorie.

De dubbelaspecttheorie zegt dat er in de wereld dingen en personen bestaan, dat dingen enkel fysische eigenschappen hebben, en dat

personen zowel fysische als mentale eigenschappen hebben. Personen hebben dus twee aspecten: het fysische en het mentale zijn twee aspecten van hetzelfde, namelijk van de persoon. Tegen deze theorie kan men verschillende bezwaren opwerpen die ik zal trachten te beantwoorden.

a) "De theorie erkent niet alleen een dualisme van eigenschappen, maar ook een dualisme van entiteiten, namelijk fysische dingen en personen" Mijn antwoord is: dat is waar, maar dat doen het token-fysicalisme en het *anomalous monism* ook. De laatste twee erkennen expliciet twee soorten eigenschappen (evt. manieren van kennen, zie 1.2) die niet tot elkaar herleidbaar zijn, maar verhullen dat zo'n dualisme van eigenschapen een dualisme van entiteiten impliceert (zie 6.4).

b) "De dubbelaspecttheorie biedt geen mogelijkheid om te verklaren hoe die fysische en mentale eigenschappen samenhangen in de persoon, en wat de grond is van die twee aspecten". Mijn antwoord is wederom dat is waar, maar dat geldt ook voor het token-fysicalisme en het *anomalous monism.* Ook beide laatste beweren dat er maar weinig brugprincipes of psycho-fysische wetten zijn, en dat er ook maar weinig van zulke principes of wetten kunnen zijn (zie 3.3 1). Ook die twee theorieën hebben de hoop opgegeven te laten zien welke fysische eigenschappen de noodzakelijke en voldoende voorwaarden vormen voor (of type-identiek zijn met) mentale eigenschappen. Voorzover er psycho-fysische wetten zijn, spreekt de dubbelaspecttheorie niet van identiteit van eigenschappen maar van correlatie van eigenschappen. Aangezien het constant samengaan van bepaalde mentale eigenschappen met bepaalde fysische eigenschappen slechts zelden voorkomt, is de voornaamste reden om van identiteit te spreken al weggevallen Bovendien ontloopt de dubbelaspecttheorie met het spreken van correlatie bij psycho-fysische wetten de problemen van de identiteitstheorie met de identificatie van eigenschappen (zie 1 2); zij erkent het dualisme van eigenschappen dat nodig is om de correlatie of de identiteit überhaupt te kunnen stellen.

c) "De dubbelaspecttheorie heeft moeite met de veroorzaking van gedrag". Mijn antwoord is· dit is inderdaad een moeilijk punt, maar misschien oplosbaar als we stellen dat de *gebeurtenissen* die met personen te maken hebben óók twee aspecten hebben, net als personen

zelf. Bepaalde gebeurtenissen worden geïdentificeerd als handelingen wanneer ze veroorzaakt worden door de redenen van personen. Die veroorzaking is dan geen *fysische* veroorzaking, en redenen kunnen net zo min als meningen of pijnen gevonden worden in de vorm van fysische dingen in de persoon. Onder een ander aspect wordt dezelfde gebeurtenis gezien als een gebeurtenis met fysische eigenschappen, veroorzaakt door de fysische eigenschappen van de persoon. De dubbelaspecttheorie kan dus toegeven, met bijvoorbeeld Fodor en Davidson, dat redenen oorzaken zijn, maar dan zijn het geen *fysische* oorzaken.

d) "Niet alleen personen hebben twee aspecten, computers en robots hebben ook twee aspecten. Maar computers en robots zijn wel degelijk fysische dingen, dus personen zijn dat ook". Mijn antwoord is: dit bezwaar verwart het *toeschrijven* van twee aspecten met het *hebben* van twee aspecten. Het bezwaar kan van de hand gewezen worden met eenzelfde argument als dat tegen Dennett (zie 5.8). zelfs al zouden alle entiteiten in de wereld fysische dingen zijn, waarvan aan sommige ook mentale eigenschappen worden *toegeschreven,* dan nog moet er tenminste één toeschrijver zijn die (misschien niet dezelfde, maar toch enige) mentale eigenschappen *heeft.* Immers, de mentale eigenschappen gaan niet samen met bepaalde fysische eigenschappen, en kunnen dus niet op grond van fysische eigenschappen worden toegeschreven. De toeschrijver moet dus fysische eigenschappen op een bepaalde manier *interpreteren,* of reageren op niet-fysische eigenschappen. In het eerste geval moet de toeschrijver kunnen interpreteren, en daarvoor moet hij zelf mentale eigenschappen hebben, en als hij niet interpreteert maar zelf op fysisch verklaarbare wijze reageert op andere systemen, moeten die andere systemen niet-fysische eigenschappen hebben.

e) "Een niet-fysicalistische filosofie leidt tot allerlei raadsels en mysteries, maar niet tot wetenschap". Mijn antwoord is: de cognitieve psychologie (en wat dat betreft, de taalwetenschap en ook alle niet-empirische wetenschappen) is in feite niet fysicalistisch en niet gebaseerd op een fysicalistische filosofie, maar het is wel een wetenschap. De empirische successen van de cognitiewetenschap, waarvan we zagen dat ze geen steun konden geven aan een fysicalistische theorie van het mentale, ondersteunen juist wel de

dubbelaspecttheorie. De cognitiewetenschap is door en door intentioneel, 'intentionaliteit' is er een primitieve notie, en er is voortdurend sprake van het hebben van meningen en wensen en doelen en het nemen van beslissingen Dat sluit uitstekend aan bij de dubbelaspecttheorie, die immers ook niet verlangt dat mentale eigenschappen gereduceerd worden tot fysische.

6.6. Gevolgen voor de cognitieve psychologie.

Als nu de token-fysicalistische theorie van het mentale niet houdbaar is, omdat intentionaliteit niet te verklaren valt in fysicalistische termen, maar de dubbelaspecttheorie van het mentale is wel houdbaar, wat zijn dan de gevolgen voor de cognitieve psychologie? De cognitieve psychologie zag zichzelf altijd als een fysicalistische theorie, gefundeerd door een fysicalistische filosofie van het mentale. Wat gebeurt er als die funderende filosofie van het mentale niet fysicalistisch is, maar een dubbelaspecttheorie? Voor de cognitieve psychologie verandert er niet zoveel. Het werk dat men doet wordt op zichzelf niet aangetast. Alleen de *opvatting* over wat men aan het doen is zou veranderd moeten worden Wat de cognitieve psychologie *niet* doet (maar wel zegt te doen), is het geven van fysicalistische verklaringen voor mentale veroorzaking. Wat ze *wel* doet (zonder dit te zeggen), is het verklaren van intelligentie en het systematiseren en verfijnen van de zogenaamde *folk-psychology*. Verklaringen als: "Hamlet vermoordde de man achter het scherm omdat hij dacht dat het zijn oom was" zijn typisch psychologische verklaringen, maar het zijn geen verklaringen in termen van fysische oorzaken. De psychologie blijft een onreduceerbare, zelfstandige wetenschap. De computerprogramma's die gemaakt worden zijn modellen om de gedragssequenties en -regelmatigheden te beschrijven en te systematiseren. Men moet alleen niet de fout maken te stellen dat de gepostuleerde toestanden en processen als *fysische* toestanden en processen in het fysisch organisme aanwezig zijn. Zolang men zich niet schuldig maakt aan deze reificatie, zolang de ontologische kwesties buiten beschouwing blijven, valt er op de cognitieve psychologie weinig aan te merken. Velen zien hun werk al op deze manier. Winograd zit

met zijn taalonderzoek op dit spoor wanneer hij zegt.

> "In saying that a representation is "present in the nervous system", we are indulging in misplaced concreteness and can easily be led into fruitless quests for the corresponding mechanisms ... there are important regularities in the "descriptive domain" that *do not* correspond to mechanisms" (Winograd 1980, 227-228.

Cummins geeft, als ik hem goed begrijp, dit spoor aan in zijn boek *The nature of psychological explanation* (1983), wanneer hij betoogt dat computermodellen niet een causaal mechanistische verklaring voor gedrag vormen, maar een formeel model om het gedrag te systematiseren, en om bepaalde capaciteiten te verklaren. De modellen laten bijvoorbeeld zien *hoe* de successieve stappen in een redenering gaan, niet *wat* het is om op grond van een bepaalde meningen tot een conclusie en tot gedrag te komen. En Searle zit met zijn analyses van taaldaden en van intentionaliteit op dit spoor, zolang hij zich onthoudt van ontologische kwesties en niet probeert een fysicalistische theorie te geven van *wat intentionaliteit is* (b.v. Searle 1983).

De cognitieve psychologie heeft veel tot stand gebracht. Ze heeft ons inzicht in de aard van de cognitie zeer vergroot. Ze heeft, voor een belangrijk deel althans, een verklaring kunnen geven voor intelligentie. Waar ze geen verklaring voor geeft, omdat ze het altijd vooronderstelt, is intentionaliteit. Als intentionele theorie biedt ze geen fysicalistische oplossing voor het lichaam-geest probleem, en is ze zelf ook geen fysicalistische theorie. Maar de cognitieve psychologie als intentionele theorie is wel compatibel met een andere oplossing voor het lichaam-geest probleem: de dubbelaspecttheorie.

SUMMARY.

Man or machine? The mind-body problem in cognitive psychology.

This study explores the possibility of a psychology dealing with beliefs and desires and feelings; with what is called in everyday usage 'the mental'. It aims at a philosophical position vis-a-vis the mind-body problem (what is the physical and what is the mental, what distinguishes them, what is the relation between them) that acknowledges, at least provisionally, our everyday experience of such mental phenomena; a philosophical position capable of grounding such a psychology.

Modern cognitive psychology is such a psychology that freely uses mentalistic terms. Cognitive psychology also claims to offer a physicalistic solution to the mind-body problem. Cognitivists find themselves strengthened in their conviction that in the final analysis people are nothing but physical systems, by the existence of 'intelligent' computers - purely physical systems the behavior of which may be explained both in physical terms and in mentalistic terms, such as beliefs and desires.

In this book the various readings of this physicalist position are reviewed, as they appear in the work of such leading philosophers as J. Fodor, D. Dennett and others. The following claims are defended:
1) that the physicalist premiss of the man-machine identity is untenable,
2) that it is nevertheless possible to accept the results of research in present day cognitive science without endorsing that premiss, and
3) that a non-physicalist position on the mind-body problem, viz. the double aspect theory, is in fact more compatible with cognitive science.

In *chapter 1* the problem is stated. There is a review of the various positions possible vis-a-vis the mind-body problem, and the physicalistic position held in cognitive psychology is outlined. It is argued in that discipline that intentionality is the mark of the mental, but that computers may also be ascribed intentionality. This physicalist position is to be elaborated in the next chapters.

In *chapter 2* the empirical evidence for the physicalist position is evaluated. The chapter deals with the question in which respect, and under which level of description, one may speak of a man-machine identity (the grain problem), and also with the question whether there are certain cognitive processes beyond the capacity of computers (the frame problem). The conclusion is that the physicalist position is not (empirically) supported by the existence of intelligent computers. A priori arguments for the physicalist position will be discussed in the next chapters.

Chapter 3 gives a general outline of functionalism, contrasted with behaviorism and with identity theory.

Chapter 4 deals with the realist reading of functionalism, according to which mental states really do exist in an organism, on the basis of an analysis of the work of J. Fodor. The fact that he postulates internal representations as explicitly present, physical (neural) entities, leads to three interrelated problems:
1) What makes those entities representations of the world?
2) What makes them represent anything at al?
3) For whom do they represent something?
Various attempts to solve these problems, by Fodor and by others, are reviewed. The conclusion is that no solutions have been found, because intentionality, as a mark of the mental, is all along being presupposed, and not explained; therefore no physicalistic theory of the mind is being offered.

Chapter 5 deals with the instrumentalist reading of functionalism, according to which mental states are only ascribed for practical purposes, but do not really exist, on the basis of an analysis of the work of D. Dennett. His arguments for eliminating many mental states are outlined, and criticized for their verificationism. The main point of criticism, however, is that Dennett's theory, even in its most extreme third person reading, still cannot explain the intentionality of the *ascriber* of beliefs and desires, as opposed to the intentionality of the *ascribee*, which is explained away The intentionality of the ascriber is presupposed; hence again no physicalistic theory of the mind is being

offered.

Chapter 6 concludes that cognitive science fails to explain intentionality, and always presupposes its existence. This presupposition of intentionality is only problematic if one insists upon a physicalistic theory of the mind. There is however no empirical evidence for physicalism. Cognitive science does not in fact support physicalism, and does not presuppose it either. A non-physicalistic theory of the mind is suggested, the double aspect theory, which is more compatible with the actual practice of cognitive science, if not with its self-conception.

1 De beschrijvingen van gedrag en gedragsdisposities die Ryle (1949) geeft zijn zeker geen beschrijvingen in *fysische* termen Wat dat betreft is Ryle geen fysicalist te noemen. Het streven van de behavioristische psychologie is wel altijd geweest om stimulus en respons in fysische termen te beschrijven.

2. Een mogelijke uitzondering op dit niet-dingachtige karakter van het eigen lichaam kan men beleven in door drugs veroorzaakte hallucinaties van 'uittreding' Men 'ziet' dan het eigen lichaam als een ding, van buitenaf.

3. Waarom doet het nieuwe Amerikaanse spel van 'dwergwerpen' (sterke mannen proberen een dwerg zo ver mogelijk van zich af te werpen) niet alleen ridicuul maar ook onaangenaam aan? De dwerg komt altijd zacht terecht en wordt goed betaald, dus daar kan het niet aan liggen. Wat onaangenaam aandoet is dat het lichaam als een materieel object wordt opgevat en geenszins, zoals dat bij judo of worstelen of boksen wel het geval is, als het lijf van een ander. Buiten een wetenschappelijke of laboratoriumcontext, waar abstractie gepast is, komt ons dat bijna obsceen voor.

4. Dit punt wordt vooral duidelijk bij de dood· bij de dood van de persoon blijft het lichaam als fysisch object over: de stoffelijke resten. Je zou je kunnen voorstellen dat zo'n lijk nog een tijd op ingenieuze wijze door electrische stimulatie bewogen kan worden. Het blijft dan evenwel een zombie, een wandelend lijk, en geen persoon. Dennett en Hofstadter (1981) spelen met soortgelijke gedachten (b v. in 'Where am I?'), maar gaan er juist wel vanuit dat het lichaam een machine is, een instrument, maar dan van de hersenen. Zij lijken te stellen dat de persoon identiek is aan de hersenen, mogelijk in aansluiting op de gelijkstelling van 'dood' met 'hersendood' in de medische praktijk van de laboratoriumgeneeskunde. Zie in dit verband ook de schitterende

roman *The mind-body problem* van Rebecca Goldstein (1983)

5. Deze problemen worden door Husserl gesignaleerd (zie b.v. Strasser 1965, De Boer 1978). Het zou interessant zijn om precies na te gaan in hoeverre mijn kritiek op Fodor's theorie, die Brentano's notie van intentionaliteit heeft overgenomen en uitgewerkt, overeenkomt met Husserl's kritiek op Brentano. Zo'n vergelijking zou in het kader van dit werk evenwel te ver voeren. Een aanzet tot de bestudering van het verband tussen Husserl en de cognitiewetenschap wordt gegeven door Dreyfus in zijn bundel *Husserl, intentionality and cognitive science* (1982).

6 Men kan zich afvragen of beide soorten mentale toestanden of processen nog wel iets met elkaar gemeen hebben. Rorty (1970b, 1980) b v meent van niet, en vindt dat de mentale toestanden met intentionaliteit zoals meningen en wensen helemaal niet een lichaam-geest probleem opleveren. En Dennett (1978b, 32) zegt heel expliciet "*... there is nothing it is like* to believe that p, desire that q, and so forth". Maar b.v. McGinn (1982a) wil beide soorten mentale toestanden verbinden door het kenmerk 'bewustzijn', en rekent onbewuste, niet expliciet gerepresenteerde meningen en wensen tot de uitzonderingen.

7. Voor het overige is Searle's opvatting van intentionaliteit afwijkend van Brentano's notie, en van de uitwerking van die notie in de cognitiewetenschap. Voor Searle bestaat het begrip 'intentionele inexistentie' niet. Voor Searle is datgene waar een propositionele attitude op gericht is juist wèl een object in de buitenwereld. Wat bij Brentano het immanente, intentionele object is is bij Searle de (propositionele) *inhoud* van de attitude. Bij Searle heeft een propositionele of psychologische attitude altijd een inhoud, maar is ze soms nergens op gericht, heeft ze soms geen object De opvattingen van Searle over intentionaliteit worden niet verder uitgewerkt in dit proefschrift.

8. In latere teksten spelt Boden 'intentionaliteit' weer wel met een t (Boden 1981)

9 Zie echter ook Fodor 1968 voor een meer filosofische uiteenzetting van de relatie mentalisme-materialisme in de psychologie.

10. Zie voor een uitgebreidere bespreking van Boden's positie Meijsing 1985.

11. In een televisie interview met Adriaan van Dis zegt Boden dat machines, gezien hun organisatie, wel emoties moeten hebben als ze ingewikkeld genoeg zijn; in die zin zijn machines dus niet anders dan machines. Maar even later zegt ze. "We are so much more than machines; we are also *emotional* beings". Zie ook Calis 1985, Bem 1985)

12 Hoewel Boden later in de richting van Dennett lijkt te gaan, bijvoorbeeld: "But in the case of a modern computer running a complex program, there is no alternative to adopting what D.C. Dennett has called 'the intentional stance' toward the machine and its program if we are to understand and explain what is going on" (Boden 1981, 3).

13. S. Silvers (persoonlijke communicatie) suggereert dat het concept verkregen zou kunnen worden door associaties tussen de lichtjes te registreren. Hij lijkt daarbij te denken aan de nieuwe trent van connectionistische netwerken (*new connectionism*). Maar als het zo simpel ging, waarom waren dan de perceptrons niet meer succesvol? Aangezien een connectionistisch netwerk formeel equivalent is met een perceptron (een finite-state machine), is het mij niet duidelijk waarom een connectionistisch netwerk beter zou werken of iets nieuws oplevert ten opzichte van de perceptrons. Overigens is het associationisme van vele connectionistische netwerken niet het enige bezwaar tegen connectionisme. Het voornaamste bezwaar is dat in een netwerk de labels van de knopen geen rol spelen; het zijn lege atomische entiteiten zonder interne structuur die volledig gedefinieerd zijn door hun positie in het netwerk. Hiertegen zijn dezelfde bezwaren in te brengen die men kan hebben tegen elke zgn. fusie-theorie volgens welke interne representaties geen interne structuur hebben (zie verder

4.2)

14 N.B. Volgens Chisholm was het *noodzakelijk* om over het psychische te spreken in intensionele zinnen, terwijl we over het fysische alles wat er te zeggen valt volledig en adequaat in niet-intensionele zinnen kunnen zeggen.

15. De geschiedenis van de computer is niet erg oud: nauwelijks veertig jaar. Toch waren er al eerder voorlopers van onze huidige computers (Goldstine 1972, New Scientist 1983). Het telraam, misschien wel het eerste rekentuig, was al vijfduizend jaar geleden in gebruik. De eerste mechanische computer werd gebouwd door Pascal, en Leibniz ontwierp een betere versie in 1673. Charles Babbage ontwierp de eerste grote computer, de 'Difference Engine', in Engeland in 1812, om mathematische functies te berekenen In 1822 begon hij er een klein model van te bouwen en de regering toonde interesse. Tien jaar later was een deel af. Maar in 1833 had hij plannen voor een 'Analytical Engine', een voorloper van de hedendaagse *general purpose* computer. Deze machine was evenwel zo ambitieus opgezet dat hij nooit gebouwd werd, al bleef Babbage er zijn hele leven aan werken. Het ontwerp noemde het gebruik van voorgeprogrammeerde sequentiële controle over rekenkundige operaties, iets waar de huidige computers nog steeds mee werken (Morrison and Morrison 1961). Ongeveer vijftig jaar later, in 1889, patenteerde Dr. Herman Hollerith de Hollerith ponskaart, die later door IBM gebruikt werd. In de dertiger jaren van deze eeuw beschreef Dr. Howard Aiken de eerste moderne voorgeprogrammeerde computer, de Mark I, die in 1944 in de Harvard universiteit gereed kwam. Deze gebruikte geponste papierband als input en had een geheugen dat bestond uit duizenden electromagnetische relays. Twee jaar later werd de ENIAC (Electronic Numerical Integrator And Computer), ontworpen door J.P. Eckert en J.W. Mauchly, gebouwd in de universiteit van Pennsylvania. De electronische computer was een feit geworden. Twee belangrijke vernieuwingen werden uitgevonden in 1945 door Dr. John von Neumann: hij gebruikte het binaire getallenstelsel in plaats van het decimale, en sloeg de programma-instructies ook in

het geheugen op De oudere machines hadden het geheugen alleen benut voor data-opslag; de programma-onderdelen werden extern ingevoerd op ponsband of -kaart, telkens wanneer een onderdeel was afgerond In 1949 werden de EDVAC (Electronic Discrete Variable Automatic Computer) en de EDSAC (Electronic Delay Storage Automatic Computer) gebouwd volgens von Neumann's principes (Aaronson et al 1976). Vanaf de vijftiger jaren is de technische ontwikkeling van de computer met sprongen vooruit gegaan (al zijn nog steeds de meeste computers van het zogenaamde von Neumann-type).

16. Waar ik artificiële intelligentie in engere zin bedoel zal ik het voluit schrijven zonder hoofdletter, voor Artificiele Intelligentie in ruime zin gebruik ik de afkorting AI of het woord voluit met hoofdletters.

17. Men kan verschillende opvattingen hebben over deze expertsystemen. Over het algemeen meent men dat deze expertsystemen een surrogaat kunnen zijn voor een persoon die hetzelfde werk doet. Er wordt dan aan het systeem intelligentie toegeschreven. Maar sommigen, die zich ook tot de cognitiewetenschap rekenen, denken er anders over Zo zegt Winograd: "From a viewpoint of human interaction we see the computer's role differently. It is not a surrogate expert, but an intermediary - a sophisticated medium of communication. A group of people (typically including both computer specialists and experts in the domain) build a program incorporating a formal representation of their beliefs. The computer communicates their statements to users of the system, typically doing some combinations and rearrangements along the way" (Winograd 1980, 235).

18. Vergelijk ook de Duyker-lezing van Margaret Boden (1985), waarin ze een *pendulumswing* in de AI laat zien van zeer 'hersenachtige' machines naar machines die helemaal niet op hersenen en neurofysiologie gebaseerd zijn en weer terug naar meer 'hersenachtige' machines.

19 Voor een scherpe veroordeling van veel te optimistische voorspellingen in de AI zie Dreyfus (1972) en (1979).

20 Vermeld in Dennett (1983/1984, 104).

21. De filosoof Popper laat zien dat Homerus als eerste robots heeft beschreven Homerus spreekt van gouden dienstmeisjes, die eruit zien als echte meisjes en hun scherpe verstand bewijzen door hun intelligente spreken en hun vaardig handelen. Popper merkt op: ". . one might perhaps find traces of Homer's reading of G Ryle" (Popper 1977, 4)

22. Het hierboven gegeven voorbeeld van een Turingtest-conversatie is op dat punt niet helemaal duidelijk. Wanneer de vragen van te voren ingeleverd zijn en men het verslag later in zijn geheel op schrift terugkrijgt, kan men de vermelde reactietijden beschouwen als berekende responsen, die niet overeen hoeven te komen met de tijd die de machine nodig heeft om het antwoord te berekenen. Het programma heeft dan een procedure die een plausibele reactietijd berekent en afdrukt vóór het eigenlijke antwoord. Maar als de conversatie *on line* plaatsvindt, dan is er sprake van echte reactietijden.

23. De kritiek van de Churchlands is een kritiek tegen een werkverdeling tussen psychologie en neurofysiologie überhaupt. Hun voorstellen bepleiten een volledig neurofysiologische psychologie, en een afschaffing van een psychologie die werkt met eigensoortige verklaringen in termen van regels en representaties. Deze voorstellen sluiten aan bij hun eliminatief materialisme.

24. Gall meende ook dat die verticale faculteiten in duidelijk onderscheidbare hersengebieden gelocaliseerd waren, die groter waren naarmate de faculteit beter ontwikkeld was. Bovendien dacht hij dat de schedel al die hobbels en knobbels in de hersenen precies volgde. Dat leidde tot de frenologie met haar schedelmetingen en haar praat over wiskundeknobbels, een leer die ook de verticale faculteiten-psychologie in discrediet heeft

gebracht

25. Wanneer de netvliesbeelden van een stilstaande kamer bewegen omdat ik mijn hoofd of ogen beweeg, dan lijkt de kamer toch gewoon stil te staan Als ik evenwel met mijn vinger tegen mijn oogbal duw, en hem zo beweeg, lijkt de kamer wel te bewegen Het input-systeem heeft niets aan de informatie, die mijn centrale zenuwstelsel wel heeft, dat ik het ben die mijn oogbal beweegt het is voor die informatie afgesloten, ondoordringbaar, ingekapseld!

26. Over de vraag naar de formaliseerbaarheid van alle kennis zijn ook verhitte discussies gevoerd die zich vooral concentreren rond de onvolledigheidsstelling van de wiskundige Gödel. Volgens deze stelling bevat ieder formeel systeem van uitspraken dat voldoende rijk is een uitspraak die waar is in dat systeem maar onbewijsbaar Aangezien een computerprogramma een formeel systeem is kan een computerprogramma dus altijd minder dan een mens, die wel de waarheid van die ene onbewijsbare uitspraak kan zien, zo luidt de redenering van Lucas (1961). De tegenwerping is dan dat mensen dit herkennen van de Godel-zin doen vanuit een ander, meer omvattend formeel systeem De dreiging van een oneindige regressie van steeds meer omvattende systemen wordt afgewend door te stellen dat mensen niet oneindig vaak kunnen 'Gödelen' (b.v. Hofstadter 1979, Hutton 1976). Ik zou daar tegenin willen brengen dat mensen inderdaad niet oneindig vaak kunnen 'Godelen' - we zijn slechts eindig en we hebben wel wat beters te doen - maar dat we, als het erop aan komt, wel onbepaald vaak kunnen 'Godelen'. Als we ermee ophouden is dat of omdat we geen zin meer hebben, of omdat we erbij neervallen, maar *niet* omdat we in ons meest omvattende formele systeem zijn aangeland.

27. Maar niet aan de hand van Dreyfus, wiens kritiek op de AI hem niet in dank is afgenomen - hij moet zowat de meest uitgescholden persoon in de cognitiewetenschap zijn. In 1966 fulmineerde S. Papert "I protest vehemently against crediting Dreyfus with any good. To state that you can accociate yourself with any of his conclusions is unprincipled" (Papert 1966, 117) En nog onlangs

beweerde een cognitiewetenschapper vreselijk boos te zijn geworden omdat ik een argument van Dreyfus gebruikte (persoonlijke communicatie)

28 Ook het probleem van Achilles en de schildpad van Lewis Carroll heeft aandacht gekregen Wanneer je alle kennis wilt formaliseren in als-dan regels krijg je een oneindige regressie van regels die expliciteren wanneer een regel van toepassing is

29 Ik kan niet nalaten een klein *persuasive* argument te geven het is toch wel opvallend dat nog in 1985 de wetenschapsbijlage van het NRC-Handelsblad als staaltjes van mooie AI-programma's kwam met voorbeelden van Weizenbaum's ELIZA (uit 1965) en Winograd's SHRDLU (uit 1971) Op geen van beide programma's is ooit voortgebouwd En is het niet opvallend dat Winograd s programma eigenlijk niet veel meer kan doen dan het wereldberoemde 'voorbeeld' dat steeds maar weer gebruikt wordt (Brandt Corstius 1981)?

30 Voor een kritiek op deze aspecten van het behaviorisme zie Koch (1959-1963, 1964)

31 Dennett vertelt dat veel Amerikaanse anesthesisten zo'n soort middel toedienen bij de anesthesie "to get us off the hook" voor het geval de patient toch iets gevoeld zou hebben

32 Dennett meent, in zijn 'A cure for the common code?' (in Dennett 1978b), dat Fodor Ryle verkeerd heeft begrepen Volgens Dennett denkt Fodor dat op de vraag 'What makes the clown's clowning clever?' slechts een antwoord mogelijk is een causaal antwoord Maar volgens mij denkt Fodor dat Ryle denkt dat er maar een antwoord mogelijk is een conceptueel antwoord Fodor wil enkel laten zien dat een causaal antwoord niet onmogelijk is.

33. De reclames voor Wheaties, waarin telkens kampioenen beweren dat ze Wheaties eten, geven in feite antwoord op de vraag in de tweede zin. Maar tevens wordt gesuggereerd dat er ook een

causaal antwoord geldig is Mogelijk blijft evenwel dat de reclame alleen wil inspelen op het *'image building'* van de Amerikanen, een aspect van reclame dat op de hak wordt genomen door de popgroep Rolling Stones in hun lied 'Satisfaction' waar ze zingen "But he can't be a man 'cause he doesn't smoke the same cigarettes as me".

34. Swart (1985) wijst het argument van Davidson tegen het behaviorisme wel op zeer vreemde wijze van de hand Hij citeert Davidson uitvoerig, en met name de zin: "Suppose, we *try* to say, not using any mental concepts, what it is for a man to believe (Davidson 1980, 217, mijn (MM) cursivering). Vervolgens zegt Swart: "I have quoted at length to bring out the full rhetorical force of the objection ... it doesn't even seem to carry much weight .. we note first of all that the objection begins with admitting that a certain mental state ('believing ...') *can* be defined behaviouristically" (Swart 1985, 148, zijn (HS) cursivering). Davidson geeft dat natuurlijk juist niet toe! (Ik citeer hier uitgebreid om de kracht van Swart's rhetoriek te laten zien.)

35. Ofschoon Dennett een artikel de titel geeft 'Why you can't make a computer that feels pain' is het niet duidelijk of hij die vraag wel wil beantwoorden, of zelfs maar stellen. Gezien zijn uitspraken dat hij voor wat betreft pijn een eliminatief materialist is, en dat hij een verificationist is, zou hij moeten zeggen dat je zo'n computer wèl kunt maken. Zijn artikel bevat een schat aan discussiemateriaal, maar ik ben er niet in geslaagd Dennett's conclusie eruit te distilleren.

36. Ik vertaal het Engelse *'believe'* voortaan met 'menen', in de zin van 'geloven', en *'belief'* met 'mening'. Het is wel niet een exacte vertaling, maar er ontstaan anders lastige problemen wanneer er gesproken wordt over *'beliefs'* in het meervoud.

37. Het onderscheid tussen de zin (Sinn) en de referentie (Bedeutung) van een teken stamt van de Duitse wiskundige en filosoof Gottlob Frege. De referentie van een teken of uitdrukking is het object

(in de wereld) dat erdoor aangeduid wordt De zin van de uitdrukking is, zoals Frege zegt, 'de manier van presentatie' van het aangeduide object. Een bekend voorbeeld om het onderscheid tussen zin en referentie duidelijk te maken is het volgende. de uitdrukkingen 'organismen met een hart' en 'organismen met nieren' hebben dezelfde referentie, maar verschillen in zin Wanneer we ons alledaagse woord 'betekenis' in deze termen willen uitdrukken, dan komt Frege's zin (Sinn) overeen met de betekenis van een uitdrukking. In de tekst zal ik verder de term 'betekenis' in verband met het betekenisprobleem gebruiken voor de Fregeaanse 'Sinn", omdat het woord 'zin' vanwege de meerduidigheid nogal lastig is. De betekenis van een uitdrukking is dus de manier van presentatie, de referentie geeft aan wat of wie we met die uitdrukking bedoelen (op dit punt zijn het Engels *meaning* en het Duitse *Bedeutung* ambigu) (Frege 1952 (1892)). Interne representaties hoeven niet altijd een referent in de buitenwereld te hebben (vgl. Brentano's 'intentionele inexistentie' en Chisholm's 'niet opgaan van existentiele generalisatie), maar ze hebben wel altijd een betekenis Let wel, ik definieer 'betekenis' hier *niet* als bepaald door de interrelatie tussen interne representaties. (Zie voor de relatie tussen Brentano's notie van intentionaliteit en Frege's onderscheid tussen betekenis en referentie Føllesdal 1969 en Stegmüller 1976)

38. Men zou kunnen zeggen dat er goede redenen zijn om DNA als een code te zien, maar is het niet zo dat alle biochemische processen direct reageren op *vorm* en niet op inhoud? De DNA-code staat niet voor iets anders, maar zet andere processen in gang.

39. Vreemd genoeg kende Fodor deze kritiek al eerder. In een artikel uit 1978, 'Tom Swift and his procedural grandmother' (herdrukt in Fodor 1981a), gebruikt Fodor hetzelfde voorbeeld dat hij zegt van Georges Rey gehoord te hebben, in een iets andere context (zie 4.6.1). Hij schijnt niet gezien te hebben dat hetzelfde argument hier tegen zijn eigen theorie gebruikt kan worden.

40. Cummins (1983, 43) beweert dat het een vergissing is te klagen

over de multi-interpreteerbaarheid van computationele systemen. Twee verschillende interpretaties (oorlog of schaken) verklaren het systeem even goed; er is geen noodzaak te kiezen tussen twee verklarende interpretaties Maar Fodor's probleem (en het mijne) is dat die multi-interpreteerbaarheid *ergens* moet stoppen. Als Fodor denkt aan RR (of ik aan GV), dan moet zijn (of mijn) interne representatie uniek geïnterpreteerd zijn als een representatie *van* RR (of GV).

41. Dreyfus (1980) valt Fodor op dit punt aan: volgens hem heeft Fodor een theorie nodig van *'as-if'* semantiek, een theorie over hoe men in ieder geval meent te refereren, ook al is de vraag over echte referentie uitgesloten. Dreyfus maakt hierbij een onderscheid tussen *Dasein 1,* een term die duidt op het werkelijk ingebed zijn in de fysische wereld, en *Dasein 2,* een term die daarentegen duidt op het ingebed zijn in een sociale wereld, een achtergrond van sociale praktijken. Fodor lijkt alleen *Dasein 1* te zien.

42. De termen in een programma hebben zelf vaak niet de bedoelde interpretatie uit de natuurlijke taal (zie 2.2.1).

43. De suggestie die hierachter zit is dat het evolutionaire proces zelf op de een of andere manier een correspondentie produceert tussen onze woorden en mentale representaties en externe dingen; we zouden niet overleven als er niet zo'n correspondentie was.

    Putnam geeft echter een argument dat ook in dat geval er wel enige *'latitude'* in het toeschrijven van een interpretatie is (Putnam 1981, 38-41). Dat argument gaat als volgt:

    Sommige van onze meningen zijn nauw verbonden met handelingen. Als teveel van onze directieve meningen onwaar zijn, zullen we teveel onsuccesvolle handelingen verrichten, dus de waarheid van onze directieve meningen is noodzakelijk voor overleving. Onze directieve meningen hangen samen met allerlei andere meningen. Dus al onze meningen moeten, tenminste bij benadering en grotendeels, waar zijn.

    Maar Putnam laat zien dat het vastleggen van de waarheidswaarde van hele zinnen niet bepaalt waar de *termen* naar

verwijzen In een technische uitwerking van een Quineaans argument (zie Quine 1960), laat hij zien dat, zelfs al is de waarheidswaarde van iedere zin van een taal vastgelegd voor iedere mogelijke wereld, de referentie van individuele termen onderbepaald blijft Dit demonstreert hij aan de hand van een voorbeeld waarbij een taal zo geinterpreteerd wordt dat in de zin "Een kat is op een mat", "kat" verwijst naar kersen en "mat" naar bomen. In deze niet-standaard interpretatie J blijft de waarheidswaarde van elke zin hetzelfde als in de standaardinterpretatie I. Dus voor evolutionair succes is een J-interpretatie even goed als een I-interpetatie. Ook de aanwezigheid van sensoren biedt hier geen uitkomst De sensoren geven wel aan dat er verschil is tussen katten en kersen, maar de mogelijkheid van beide interpretaties blijft bestaan. Er blijft wel degelijk *latitude*, onbepaaldheid in de interpretatie Maar die onbepaaldheid is niet specifiek voor robots En, zoals Pylyshyn opmerkt, het zou pervers zijn om in zo'n geval een niet-standaard interpretatie, of zelfs geen interpretatie, toe te kennen aan de interne representaties van de robot.

44 Nauw verwant hiermee is het aloude probleem van de stimulus-equivalentie. Onder welke fysische specificatie zijn, bijvoorbeeld, stoelen equivalent? Wat is de juiste fysische beschrijving voor poppen, potloden, po's? Ook dit probleem wordt algemeen onoplosbaar geacht. We kunnen van de meeste 'dingen' geen *fysische* kenmerken-lijst geven omdat het altijd kan voorkomen dat iets alle kenmerken heeft maar toch het 'ding' niet is, of het 'ding' is maar niet alle kenmerken heeft. De behavioristen hebben altijd gestreefd naar een fysische specificatie van de stimulus. Maar in de praktijk kwam het er zelden of nooit van. Zo geeft Guthrie in zijn laatste artikel heel expliciet de hoop op "... we find ourselves inevitably describing stimuli in perceptual terms .. it is .. necessary that they have meaning for the responding organism" (Guthrie 1959, 165, geciteerd in Koch 1964).

45. Veel hangt overigens af van wat Fodor precies bedoelt met 'fenomenologisch waarneembaar'. Soms lijkt het alsof alle

eigenschappen, zoals kleur, omvang, textuur, fenomenologisch waarneembaar zijn en alleen moleculaire en sub-moleculaire structuur niet. Dan zijn de modulen fenomenalistisch; zij zetten gewaarwordingsuitspraken om in eigenschappen van dingen. Soms ook noemt hij echter *schijnbare lengte* een fenomenologisch waarneembare eigenschap. Dat lijkt een gewaarwordings-eigenschap. In dat geval zijn de denkprocessen fenomenalistisch, omdat die van de modulen alleen een soort gewaarwordingsuitspraken, zoals schijnbare lengte, krijgen. In ieder geval heeft Fodor een theorie nodig die een principieel onderscheid kan maken tussen fenomenologisch waarneembare en niet waarneembare eigenschappen, en een principiële notie van 'hoe dingen eruit zien' (zie ook Fodor 1984b). Dat zijn op zijn zachtst gezegd uiterst problematische behoeften. Is iets wat ik alleen met een bril op kan zien een fenomenologisch waarneembare eigenschap? En alleen door een gewoon vergrootglas? Een microscoop? Waar ligt een principiële grens? Of is een bril alleen een hulpmiddel om normaal te kunnen zien, een correctie op abnormaal gezichtsvermogen, maar een vergrootglas, laat staan een microscoop niet? Wie maakt dan uit wat normaal is? (Zie ook 4.5.4.)

46. Overigens valt deze poging moeilijk te rijmen met Fodor's artikel 'The present status of the innateness controversy' (herdrukt in Fodor 1981a), waar hij zwaarwegende argumenten geeft tegen iedere mogelijkheid van reducerende definities in een taal. Hij argumenteert daar dat erg veel concepten simpel, niet samengesteld, en dus niet definieerbaar in andere termen zijn. Maar dat zou betekenen dat erg veel van de interne taal niet gedefinieerd kan worden in andere termen, en geen interpretatie kan krijgen via een combinatie van fenomenologisch waarneembare eigenschappen.

47. In dit verband is het aardig Fodor en Pylyshyn's (1981) kritiek op Gibson's perceptietheorie te vermelden. Volgens Fodor en Pylyshyn *moet* perceptie een proces zijn waarbij computaties worden uitgevoerd op fysische input. Zij zitten dan met het probleem

welke stimulus-eigenschap, fysisch gespecificeerd, (bijvoorbeeld) 'eetbaar' is, waarbij 'eetbaar' intern gerepresenteerd is Volgens de Gibsoniaanse perceptieleer is er geen sprake van computaties en representaties. Eetbare dingen zijn er gewoon in de werkelijke wereld, en het organisme kan ze (doorgaans) herkennen. Volgens Fodor en Pylyshyn is dit geen oplossing, maar een trivialisering van hun probleem.

Men kan zich evenwel afvragen *waarom* Fodor en Pylyshyn volhouden dat perceptie op een bepaalde manier *moet* gaan, als die manier leidt tot onoplosbare problemen. Voor een theorie die aanneemt dat perceptie (kennelijk) niet op die bepaalde manier gaat doen zich die problemen (wellicht) niet voor, van die theorie mag dan ook niet verwacht worden dat ze die problemen oplost

47a. In een lezing, gehouden te Nijmegen op 19 juni 1986 (na voltooiing van dit manuscript) nam Fodor weer een ietwat ander standpunt in ten aanzien van de causale theorie van representaties. Hij ziet in dat verhaal, getiteld 'How do thoughts have contents', niet langer een oplossing voor het probleem van misrepresentaties in een vorm van natuurlijke teleologie, mede om dezelfde redenen die ik genoemd heb. Hij noemt tevens als argument tegen zo'n teleologische oplossing dat een vorm van Freudiaanse repressie optimaal kan zijn voor het functioneren van een organisme, terwijl een dergelijk optimaal functioneren juist niet leidt tot de waarheid van de interne representaties.

Zijn nieuwe oplossing voor het probleem van misrepresentatie kan volgens mij de causale theorie al evenmin redden. Volgens Fodor kun je wel volhouden dat een bepaalde interne structuur (b.v. A) de semantische inhoud 'rood' heeft en niet 'rood of wit', ook al wordt A niet alleen door rode voorwerpen veroorzaakt, maar soms ook door een wit voorwerp. Dat is volgens hem omdat er sprake is van een asymmetrische afhankelijkheid De misrepresentatie van zo'n wit voorwerp als rood door A kan alleen maar vóórkomen als A ook door rode voorwerpen wordt veroorzaakt. Maar andersom kan een representatie van een rood voorwerp als rood door A vóórkomen als A nooit door witte voorwerpen wordt veroorzaakt. Of anders gezegd in alle mogelijke werelden wordt A door rode

voorwerpen veroorzaakt, en alleen in deze wereld ook wel eens door witte. Zo is misrepresentatie afhankelijk van representatie, maar niet andersom

Mijn kritiek hierop is dat de notie van asymmetrische afhankelijkheid pas gaat gelden *nadat* is vastgesteld dat A als semantische inhoud 'rood' heeft en *niet* rood of wit'. Die afhankelijkheid kan dus niet *gebruikt* worden om vast te stellen wat de semantische inhoud is

Overigens zei Fodor in de dicussie na zijn lezing dat inderdaad *eigenschappen,* en niet *entiteiten,* de oorzaak zijn van interne representaties. Er is volgens hem een nomologische relatie tussen soorten van eigenschappen en soorten van symbolen in het Mentalees Zo is er een nomologische relatie tussen de eigenschap een eenhoorn te zijn, en het symbool 'eenhoorn' in het Mentalees, *ook al is die eigenschap in deze wereld nergens geïnstantieerd.* Maar dan vraag ik mij af: wat is er gebeurd met Fodor's doelstelling om de semantische relatie uit te drukken in naturalistische (fysicalistische) termen? Eerst leek het er nog op dat volgens Fodor een causale relatie tussen een instantiatie van een eigenschap en een interne structuur de semantische interpretatie van die structuur moest vastleggen. Nu blijkt dat die causale relatie er in feite helemaal niet hoeft te zijn om die semantische interpretatie vast te leggen Kennelijk bestaat die semantische interpretatie onafhankelijk van die causale relatie. Volgens Fodor is de semantische relatie een wetmatige relatie tussen typen van eigenschappen en typen van symbolen in het Mentalees, en *niet* een causale relatie tussen individuele instantiaties van eigenschappen en individuele instantiaties van symbolen Maar daarmee zegt hij niets anders dan wat *alle* semantische theorieen zeggen. er is een relatie tussen de eigenschap een stoel te zijn en het symbool 'stoel', een relatie die Fodor eerder onïnteressant noemde. In feite heeft hij hiermee de causale theorie van representatie zover uitgehold dat het geen causale theorie van representatie meer is: Fodor is weer terug bij af.

48. Men kan zich afvragen of een combinatie van een functionele-rol-

semantiek met een of andere referentiele semantiek überhaupt mogelijk is (b.v. Silvers 1986), gezien de verschillende geaardheid (holistisch vs. individualistisch) van beide soorten semantiek. Maar zo'n vraag betreft een *algemeen* semantisch probleem; zo'n probleem mag niet opgeworpen worden voor het probleem van de semantische interpretatie van interne representaties *in het bijzonder*. Men mag uit de eventuele onoplosbaarheid van dit probleem niet concluderen dat de interne representaties geen inhoud hebben, tenzij men bereid is om te zeggen dat de natuurlijke taal ook geen inhoud heeft Waar het mij om gaat is of er een fysicalistische theorie mogelijk is voor de semantische interpretatie van interne representaties, omdat zo'n theorie nodig is voor een fysicalistische theorie van het mentale à la Fodor. De discussie rond de mogelijkheid van een (fysicalistische) semantiek überhaupt gaat het kader van dit werk te buiten

49. Dreyfus (1979, 66) gebruikt dit citaat, maar om een ander argument te illustreren, namelijk dat menselijke intelligentie niet volledig en expliciet begrepen kan worden.

50 Dennett's spelling is 'intentionaliteit' met een t Dennett doelt met zijn term volgens zijn eigen zeggen zowel op Brentano's notie van intentionaliteit als op intensionaliteit als (Chisholmse) eigenschap van linguistische entiteiten die bepaalde logische kenmerken hebben (zie 1.4.2).

51. Een argumentatie dat Fodor's theorie last heeft van homunculi, zoals ik in 4 5.3 en 4.7 heb willen aantonen, ben ik bij Dennett nergens helemaal expliciet tegengekomen. Duidelijke hints in die richting zijn evenwel te vinden in 'A cure for the common code?' (1977, herdrukt in Dennett 1978b) en in 'Styles of mental representation' (Dennett 1982-1983)

52 Heel erg 'zoiets' Dennett laat in 'Skinner skinned' (herdrukt in Dennett 1978b) zien dat dit inzicht van Skinner schier onontwarbaar vermengd is met allerlei andere bezwaren tegen mentalistisch taalgebruik.

53. Dennett beschrijft dit onschadelijk maken van homunculi in een artikel 'Artificial intelligence as philosophy and as psychology' (herdrukt in Dennett 1978b) uit 1978 Fodor beschrijft deze reeks van steeds stommere homunculi al in een artikel uit 1968 'The appeal to tacit knowledge' (herdrukt in Fodor 1981a), dat Dennett overigens wel vermeldt. De problemen die Fodor later krijgt in verband met de procedurele semantiek meent Dennett te kunnen omzeilen met zijn noties van *tacit* representatie en intentioneel systeem.

54. Dennett gebruikt, zoals veel Angelsaksische filosofen en cognitiewetenschappers, de term 'fenomenologie' voor een soort beschrijvende, op introspectie berustende psychologie

55. Churchland is er, in zijn *Scientific realism and the plasticity of mind* (1979), óók van overtuigd dat er niets gegeven is, maar dat alle observatie volledig theoriegeladen is. Desalniettemin is hij, in tegenstelling tot Rorty, wel een realist, en is hij ervan overtuigd dat een (ideaal voltooide) fysische wetenschap de enige ware beschrijving van de wereld vormt Ik vraag mij af op grond van wat hij nog kan beweren dat zo'n fysica waar is· wat zijn de gegevens die zo'n theorie moet verklaren? Bovendien is het onduidelijk hoe de *inhoud* van een onware theorie - toch zeker geen fysische entiteit, maar hooguit een mentale, een reeks meningen - zo'n reële invloed kan hebben op het, in een fysische theorie fysische, proces van waarneming (zie ook Fodor 1984b) Volgens de eliminatief materialist wordt onze waarneming dus niet bepaald door hoe de wereld is, maar door hoe we menen dat de wereld is, door hoe we de wereld voor onszelf representeren Hoe kan hij dan tegelijkertijd beweren dat meningen en representaties en intentionaliteit niet bestaan?

56 Zie ook de kritiek van Straus op de psycholoog Hebb· de schrijver Hebb die gelezen wil worden en argumenteert kan nooit het behavioristische reflex-wezen zijn dat de mens volgens de psycholoog Hebb is (Straus 1956).

57. Rorty lijkt er even in te vallen, maar weet de kuil echter keurig te omzeilen, in zijn artikel over het elimineren van het concept 'sensaties' (zie Rorty 1971a) Ik heb nergens kunnen ontdekken of hij gezien heeft dat Dennett er wel ingevallen is.

58 Overigens is het duidelijk dat beiden hier eerder over intensionaliteit-met-een-s dan intentionaliteit-met-een-t hebben; maar het is misschien intensionaliteit van zinnen als signaal voor gepostuleerde intentionaliteit in de te verklaren systemen.

59. Coulter ziet ook problemen voor de individualisatie van gedrag. Het causale netwerk in de hersenen heeft als uiteindelijk effect een bepaalde beweging. Maar het is dan nog niet duidelijk welk (aspect van) gedrag zo'n beweging vormt.

60 We hebben boven gezien dat het fysicalisme zijn ontologische 'beet' verliest als de brugformules uitzonderingen hebben. De disjunctie moet dus in alle mogelijkheden voorzien.

61. Merkwaardig genoeg noemt Fodor een soortgelijk probleem in een ander artikel, dat in 1972, dus twee jaar eerder is verschenen dan zijn artikel 'Special sciences' In 'What psychological states are not', geschreven samen met Ned Block (1972, herdrukt in Fodor 1981a), geeft Fodor het volgende voorbeeld. Stel een psychologisch predikaat p1 is coëxtensief met disjunctief fysisch predikaat A, en psychologisch predikaat p2 is coextensief met disjunctief fysisch predikaat B. Stel verder dat Si een bepaalde fysische toestand is die soms p1 heeft gerealiseerd maar niet p2, en soms p2 maar niet p1. Si is dan een disjunct van A èn van B. Maar de disjuncten van A zijn ieder afzonderlijk voldoende voorwaarde voor p1 en de disjuncten van B zijn ieder afzonderlijk voldoende voorwaarde voor p2 Daaruit volgt dat een organisme in toestand Si zowel in p1 àls in p2 is, en dat is tegen de aanname Fodor en Block zeggen: "Of course, one could circumvent this objection by including spatiotemporal designators in the specification of the disjuncts mentioned in A and B. But to do so would be totally to abandon

the project of expressing psycho-behavioral (or psycho-physical) correlations by lawlike biconditionals" (Fodor 1981a, 83). In zijn latere artikel vindt Fodor het project kennelijk niet hopeloos.

Aaronson, D., Grupsmith, E. and Aaronson, M. (1976). 'The impact of computers on cognitive psychology', *Behavior Research Methods and Instrumentation*, 8 (2), 129-138.

Anderson, J.R. and Bower, G.H. (1973). *Human associative memory*, Washington D.C., Winston.

Armer, P. (1963). 'Attitudes toward intelligent machines', in Feigenbaum, E.A. and Feldman, J. (eds), *Computers and thought*, New York, McGraw-Hill.

Armstrong, D.M. (1968). *A materialist theory of the mind*, London, Routledge & Kegan Paul.

Armstrong, D.M. and Malcolm, N. (1984). *Consciousness and causality*, Oxford, Basil Blackwell.

Austin, J.L. (1967). 'Pretending', in Gustafson, D.F. (ed), *Essays in philosophical psychology*, London, Macmillan.

Bem, S. (1985). 'Mensen en computers. Calis doet Boden geen recht', *De Psycholoog*, XX, 469-470.

Bernstein, R.J. (1971). 'The challenge of scientific materialism', in Rosenthal, D.M. (ed), *Materialism and the mind-body problem*, Englewood Cliffs N.J., Prentice-Hall.

Block, N. (1978). 'Troubles with functionalism', in Savage, C.W. (ed), *Perception and cognition. Issues in the foundations of psychology. Minnesota studies in the philosophy of science, vol. 9*, Minneapolis, University of Minnesota Press.

Block, N. (ed) (1980). *Readings in the philosophy of psychology*, London, Methuen.

Boden, M.A. (1972). *Purposive explanation in psychology*, Cambridge Mass., Harvard University Press.

Boden, M.A. (1977). *Artificial intelligence and natural man*, Hassocks, Harvester Press.

Boden, M.A. (1979). 'The computational metaphor in psychology', in Bolton, N. (ed), *Philosophical problems in psychology*, London, Methuen.

Boden, M.A. (1981). *Minds and mechanisms: philosophical psychology and computational models*, Ithaca, Cornell University Press.

Boden, M.A. (1983) 'Artificial intelligence and animal psychology', *New Ideas in Psychology,* 1 (1), 11-33

Boden, M.A. (1985). *Computer models of the mind: are they socially pernicious?* Deventer, Van Loghum Slaterus.

Boer, Th. de (1978). *The development of Husserl's thought,* Den Haag, Nijhoff.

Borst, C.V. (ed). (1970). *The mind/brain identity theory,* London, Macmillan.

Brandt Corstius, H. (1981). 'Weg met de computer!', *Kennis en Methode,* 1, 24-31.

Burge, T. (1979). 'Individualism and the mental', in French, P.A., Uehling, T.E. and Wettstein, H.K. (eds), *Midwest studies in philosophy vol. IV: Studies in metaphysics,* Minneapolis, University of Minnesota Press.

Burge, T. (1982). 'Other bodies', in Woodfield, A. (ed), *Thought and object,* Oxford, University Press.

Calis, G. (1985). 'Margaret, wat zijt gij? Bodens Duykerlezing ontwijkt de hamvraag', *De Psycholoog,* XX (9), 404-408.

Carnap, R. (1947). *Meaning and necessity,* Chicago, University of Chicago Press.

Chisholm, R.M. (1957). *Perceiving,* Ithaca, Cornell University Press.

Chisholm, R.M. (1967). 'Intentionality', in Edwards, P. (ed), *The encyclopedia of philosophy,* New York, Macmillan & Free Press.

Chomsky, N. (1959). 'A review of B.F. Skinner's *Verbal behavior', Language,* 35, 1, 26-58.

Chomsky, N. (1980). 'Rules and representations', *Behavioral and Brain Sciences,* 3, 1-60.

Churchland, P.M. (1979). *Scientific realism and the plasticity of mind,* Cambridge, Cambridge University Press.

Churchland, P.M. (1980). 'PLasticity: conceptual and neuronal', *Behavioral and Brain Sciences,* 3, 133-134.

Churchland, P.M. (1984). *Matter and consciousness,* Cambridge Mass., M.I.T. Press.

Churchland, P.S. (1980). 'Neuroscience and psychology: should the labor be divided?', *Behavioral and Brain Sciences,* 3, 133.

Colby, K.M. (1981). 'Modelling a paranoid mind', *Behavioral and Brain Sciences,* 4 (4), 515-560.

Coulter, J. (1982). 'Theoretical problems of cognitive science', *Inquiry*, 25, 3-26.

Cummins, R. (1983). *The nature of psychological explanation*, Cambridge Mass., M.I.T. Press.

Davidson, D. (1969). 'The individuation of events', in Davidson, D. (ed), *Essays on actions and events*, Oxford, Clarendon Press.

Davidson, D. (1970). 'Mental events', in Davidson, D. (ed), *Essays on actions and events*, Oxford, Clarendon Press.

Davidson, D. (1980). *Essays on actions and events*, Oxford, Clarendon Press.

Dekkers, W.J.M. (1985). *Het bezielde lichaam*, Zeist, Kerckebosch.

Dennett, D.C. (1969). *Content and consciousness*, London, Routledge & Kegan Paul.

Dennett, D.C. (1973). 'The philosophical lexicon', Oxford, manuscript.

Dennett, D.C. (1978a). 'Current issues in philosophy of mind', *American Philosophical Quarterly*, 15 (4), 249-261.

Dennett, D.C. (1987b). *Brainstorms. Philosophical essays on mind and psychology*, Brighton, Harvester Press.

Dennett, D.C. (1982a). 'Beyond belief', in Woodfield, A. (ed), *Thought and object*, Oxford, Oxford University Press.

Dennett, D.C. (1982b). 'Comment on Rorty', *Synthese*, 53, 349-356.

Dennett, D.C. (1982-83). 'Styles of mental representation', *Proceedings of the Aristotelian Society*, 83, 213-226.

Dennett, D.C. (1983-84). 'Kunnen machines denken?', *Wijsgerig Perspectief*, 24 (4), 98-108.

Dennett, D.C. (1984). *Elbow room: the varieties of free will worth wanting*, Oxford, Oxford University Press.

Donnellan, K. (1971). 'Reference and definite description', in Steinberg, D.D. and Jakobovits, L.A. (eds), *Semantics. An interdisciplinary reader in philosophy, linguistics and psychology*, Cambridge, University Press.

Dretske, F.J. (1981). *Knowledge and the flow of information*, Cambridge Mass., M.I.T. Press.

Dreyfus, H.L. (1965). 'Alchemy and artificial intelligence', *The RAND Corporation*, december 1965.

Dreyfus, H.L. (1972). *What computers can't do. A critique of artificial reason*, New York, Harper & Row.

Dreyfus, H.L. (1978). 'Empirical evidence for a pessimistic prognosis for cognitive science', *Behavioral and Brain Sciences*, 1, 105.

Dreyfus, H.L. (1979). *What computers can't do. The limits of artificial reason*, New York, Harper & Row.

Dreyfus, H.L. (1980). 'Dasein's revenge: methodological solipsism as an unsuccessful escape strategy in psychology', *Behavioral and Brain Sciences*, 3, 78-79.

Dreyfus, H.L. (ed), (1982). *Husserl, intentionality and cognitive science*, Cambridge Mass., M.I.T. Press.

Ernst, G.W. and Newell, A. (1967). *Generality and GPS*, Carnegie Institute of Technology.

Feigl, H. (1958). 'The "mental" and the "physical"', in Feigl, H. and Maxwell, G. (eds), *Minnesota studies in the philosophy of science, vol. II*, Minneapolis, University of Minnesota Press.

Feyerabend, P. (1970). 'Comment: 'Mental events and the brain'', in Borst, C.V. (ed), *The mind-brain identity theory*, London, Macmillan.

Field, H. (1972). 'Tarski's theory of truth', *Journal of Philosophy*, LXIX, 347-375.

Fink, D.G. (1966). *Computers and the human mind: an introduction to artificial intelligence*, Garden City N.Y., Anchor Books.

Fodor, J.A. (1968). *Psychological explanation. An introduction to the philosophy of psychology*, New York, Random House.

Fodor, J.A. (1975). *The language of thought*, New York, Thomas Y. Crowell.

Fodor, J.A. (1978). 'Computation and reduction', in Savage, C.W. (ed), *Perception and cognition. Issues in the foundations of psychology. Minnesota studies in the philosophy of science, vol. 9*, Minneapolis, University of Minnesota Press.

Fodor, J.A. (1980a). 'Methodological solipsism considered as a research strategy in cognitive psychology', *Behavioral and Brain Sciences*, 3, 63-109.

Fodor, J.A. (1980b). 'Searle on what brains can do', *Behavioral and Brain Sciences*, 3, 431-432.

Fodor, J.A. (1981a). *Representations: philosophical essays on the foundations of cognitive science*, Cambridge Mass., M.I.T. Press.

Fodor, J.A. (1981b). 'The mind-body problem', *Scientific American*,

244 (1), 114-123.

Fodor, J.A. (1983). *The modularity of mind,* Cambridge Mass., M.I.T. Press.

Fodor, J.A. (1984a). 'Semantics, Wisconsin style', *Synthese,* 59 (3), 231-250.

Fodor, J.A. (1984b). 'Observation reconsidered', *Philosophy of Science,* 51, 23-43.

Fodor, J.A. (1985). 'Fodor's guide to mental representations: the intelligent auntie's vade-mecum', *Mind,* XCIV (373), 76-100.

Fodor, J.A. (?). 'Narrow content and meaning holism', manuscript, Cambridge Mass., M.I.T.

Fodor, J.A. (1986). 'How do thoughts have contents?', voordracht Psychologisch Laboratorium KU Nijmegen, 19 juni 1986.

Fodor, J.A. and Pylyshyn, Z.W. (1981). 'How direct is visual perception? Some reflections on Gibson's 'Ecological Approach'', *Cognition,* 9, 139-196.

Føllesdal, D. (1969). 'Husserl's notion of a noema', *Journal of Philosophy,* LXVI, 680-687.

Frege, G. (1952 (1892)). 'On sense and reference (Über Sinn und Bedeutung)', in Geach, P. and Black, M. (eds), *Translations from the philosophical writings of Gottlob Frege,* Oxford, University Press.

George, F.H. (1962). *The brain as a computer,* London, Pergamom.

Goldstein, R. (1983). *The mind-body problem,* New York, Dell Publishing Co., Inc.

Goldstine, H. (1972). *The computer from Pascal to von Neumann,* Princeton N.J., Princeton University Press.

Gunderson, K. (1971). *Mentality and machines,* New York, Doubleday-Anchor.

Haugeland, J. (1978a). 'Nature and plausibility of cognitivism', *Behavioral and Brain Sciences,* 2, 215-226.

Haugeland, J. (1978b). 'The problem of generality', *Behavioral and Brain Sciences,* 1, 107-108.

Haugeland, J. (1981). *Mind design: Philosophy, psychology and artificial intelligence,* Cambridge Mass., M.I.T. Press.

Hayes, P.J. (1978). 'Cognitivism as a paradigm', *Behavioral and Brain Sciences,* 2, 238-239.

Heerden, J van en Zeytveld, F. van (1985). 'Kleine mens in grote mens: ecce homunculus', *Gedrag*, 13 (1), 30-48.

Hofstadter, D.R. (1979). *Gödel, Escher, Bach: an eternal golden braid*, New York, Basic Books.

Hutton, A. (1976). 'This Gödel is killing me', *Philosophia*, 6, 135-144.

Johnson-Laird, P.N. (1977). 'Procedural semantics', *Cognition*, 5 (3), 189-214.

Johnson-Laird, P.N. (1980). 'Mental models in cognitive science', *Cognitive Science*, 4 (1), 71-115.

Kalke, W. (1969). 'What is wrong with Fodor and Putnam's functionalism?', *Nous*, III (1), 83-94.

Kempen, G.A.M. (1983). 'Het artificiële intelligentieparadigma. Ervaringen met een nieuwe methodologie voor cognitief-psychologisch onderzoek', in Raaijmakers, J., Hudson, P. en Wertheim, J. (red), *Metatheoretische aspecten van de psychonomie*, Deventer, Van Loghum Slaterus.

Kempen, G.A.M. (1984). 'Inleiding', in Kempen, G A.M. en Sprangers, Ch. (red), *Kennis, mens en computer*, Lisse, Swets & Zeitlinger.

Koch, S. (ed), (1959-63). *Psychology: a study of a science*, New York, McGraw-Hill.

Koch, S. (1964). 'Psychology and emerging conceptions of knowledge as unitary', in Wann, T.W (ed), *Behaviorism and phenomenology*, Chicago, University of Chicago Press

Kripke, S. (1972). 'Naming and necessity', in Davidson, D. and Harman, G. (eds), *Semantics of natural language*, Dordrecht, Reidel.

Kuhn, T. (1962). *The structure of scientific revolutions*, Chicago, University of Chicago Press.

Lakatos, I. (1970). 'Falsification and the methodology of scientific research programmes', in Lakatos, I. and Musgrave, A. (eds), *Criticism and the growth of knowledge*, Cambridge, University Press.

Leseman, P. (1983). 'Van wat is de AI een metafoor?', *Spiegeloog*, 11 (108), 9-12.

Loar, B.F. (1980). 'Syntax, functional semantics and referential semantics', *Behavioral and Brain Sciences*, 3, 89-90.

Loar, B.F (1981). *Mind and meaning*, Cambridge, University Press.

Lucas, J.R. (1961). 'Minds, machines and Godel', *Philosophy*, 36, 112-127.

Margenau, H. (1977) 'Effects of an observer on measurements in quantum mechanics', manuscript, Yale University.

Margolis, J. (1980) 'The trouble with homunculus theories', *Philosophy of Science*, 47, 224-259.

Marr, D. (1982). *Vision*, San Francisco, Freeman Press.

Marr, D. and Poggio, T. (1977). 'From understanding computation to understanding neural circuitry', *Neurosciences Research Progress Bulletin*, 15, 470-488.

Marres, R. (1985). *Filosofie van de geest*, Muiderberg, Coutinho.

Matthews, R. (1984). 'Troubles with representationalism', *Social Research*, 51, 4, 1065-1097.

McDermott, D. (1976). 'Artificial intelligence meets natural stupidity', *Sigart Newsletter*, 57, april 1976, 4-9, herdrukt in Haugeland, D. (ed), *Mind design: Philosophy, psychology and artificial intelligence*, Cambridge Mass., M.I.T. Press, 1981.

McGinn, C. (1982a). *The character of mind*, Oxford, University Press.

McGinn, C. (1982b). 'The structure of content', in Woodfield, A. (ed), *Thought and object*, Oxford, Clarendon Press.

Meijsing, M. (1981). 'Verder lezen of niet?', in Segers, R.T. (red), *Lezen en laten lezen*, 's Gravenhage, Martinus Nijhoff.

Meijsing, M. (1984) 'Kunnen computers echt denken?', in Kempen, G.A M. en Sprangers, Ch. (eds), *Kennis, mens en computer*, Lisse, Swets & Zeitlinger.

Meijsing, M. (1985). 'Wonderen en wetenschappelijke verklaring. Een kijk op het werk van Margaret Boden', *De Psycholoog* XX, 2, 61-68.

Michie, D. (1982). 'Mind-like capabilities in computers - a note on computer-induction', *Cognition*, 12, 1, 97-108.

Miller, G.A. (1956). 'The magical number seven plus or minus two', *Psychological Review*, 63, 81-96.

Miller, G.A., Galanter, E. and Pribram, K.H. (1960). *Plans and the structure of behavior*, New York, Holt, Rinehart & Winston.

Miller, G. and Johnson-Laird, P. (1976). *Language and perception*,

Cambridge Mass , Harvard University Press.

Miller, J (1983) *States of mind*, London, British Broadcasting Corporation.

Minsky, M.L (1963). 'Steps toward artificial intelligence', in Feigenbaum, E.A. and Feldman, J. (eds), *Computers and thought*, New York, McGraw-Hill.

Minsky, M.L. (1975a). 'Frame-system theory', in Schank, R.C. and Nash-Webber, B.L. (eds), *Theoretical issues in natural language processing*, Preprints of a conference at M.I.T., june 1975

Minsky, M.L. (1975b). 'A framework for representing knowledge', in Winston, P. (ed), *The psychology of computer vision*, New York, McGraw Hill.

Minsky, M.L. and Papert, S. (1969). *Perceptrons*, Cambridge Mass., M.I T. Press.

Morrison, P and Morrison, E. (1961). *Charles Babbage and his calculating engines*, New York, Dover.

Nagel, T. (1974). 'What is it like to be a bat?', *Philosophical Review*, 83 (4), 435-450.

Neisser, U. (1967) *Cognitive psychology*, New York, Appleton Century Crofts.

Nelson, R.J. (1976). 'Mechanism, functionalism and the identity theory', *Journal of Philosophy*, LXXIII (13), 365-385.

Neumann, J. von (1956). 'The general and logical theory of automata', in Newman, J.R. (ed), *The world of mathematics*, New York, Simon & Schuster.

New Scientist (1983). 'Babbage and the birth of the computer', *New Scientist*, 99 (1375).

Newell, A. (1969). 'Heuristic programming III- Structured problems', in Aranofsky, J S. (ed), *Progress in operations research*, New York, Wiley.

Newell, A., Shaw, J.C. and Simon, H.A. (1957). 'Empirical explorations of the Logic Theory Machine. A case study in heuristics', *The RAND Corporation*, march 1957, Report P-951

Newell, A., Shaw, J.C. and Simon, H.A. (1960). 'Report on a general problem-solving program', *Proceedings of the International Conference on Information Processing*, Paris, Unesco.

Newell, A., Shaw, J.C. and Simon, H.A. (1963) 'Chess-playing

programs and the problem of complexity', in Feigenbaum, E.A. and Feldman, J. (eds), *Computers and thought*, New York, McGraw-Hill.

Nisbett, R.E. and Ross, L. (1980). *Human inference; strategies and shortcomings of social judgment*, Englewood Cliffs, N.J., Prentice-Hall.

Norman, D.A. (1980). 'Fodor's solipsism: don't look a gift horse in the ...', *Behavioral and Brain Sciences*, 3, 90.

Oettinger, A.G. (1963). 'The state of the art of automatic language translation: an appraisal', in Marchl, H. (ed), *Beiträge zur Sprachkunde und Informationsbearbeitung*, vol. I, Heft 2, München, Oldenbourg Verlag.

Papert, S. (1966) 9th RAND Symposium (november 7, 1966), 116.

Penelhum, T. (1967). 'The logic of pleasure', in Gustafson, D.F. (ed), *Essays in philosophical psychology*, London, MacMillan.

Place, U.T. (1956) 'Is consciousness a brain process?', *British Journal of Psychology*, 47, 44-50.

Place, U.T (1967). 'The concept of heed', in Gustafson, D F. (ed), *Essays in philosophical psychology*, London, MacMillan.

Polya, G. (1954). *How to solve it*, Princeton N.J., Princeton University Press.

Popper, K R. and Eccles, J C. (1977) *The self and its brain*, Berlin, Springer.

Putnam, H. (1960). 'Minds and machines', in Hook, S. (ed), *Dimensions of mind: a symposium*, New York, New York University Press.

Putnam, H. (1964). 'Robots: machines or artificially created life?', *Journal of Philosophy*, LXI, 668-690.

Putnam, H. (1965). 'Psychological predicates', in Capitan, W.H. and Merrill, D.D. (eds), *Art, mind and religion*, Pittsburgh, University of Pittsburgh Press.

Putnam, H. (1967). 'The mental life of some machines', in Castaneda, H. (ed), *Intentionality, minds and perception*, Detroit, Wayne State University Press.

Putnam, H. (1973). 'Reductionism and the nature of psychology', *Cognition*, 2 (1), 131-146.

Putnam, H. (1975). *Mind, language and reality. Philosophical papers*

volume 2, Cambridge, Cambridge University Press

Putnam, H. (1981). *Reason, truth and history*, Cambridge, Cambridge University Press

Putnam, H. (1983). 'Computational psychology and interpretation', in Putnam, H *Realism and reason. Philosophical papers vol 3*, Cambridge, University Press

Pylyshyn, Z.W (1973). 'What the mind's eye tells the mind's brain· a critique of mental imagery', *Psychological Bulletin*, 80, 1-14.

Pylyshyn, Z.W. (1978). 'Computational models and empirical constraints', *Behavioral and Brain Sciences*, 1, 93-127

Pylyshyn, Z.W. (1979). 'The rate of 'mental rotation' of images. a test of a holistic analogue hypothesis', *Memory and Cognition*, 7, 19-28.

Pylyshyn, Z.W. (1980). 'Computation and cognition: issues in the foundations of cognitive science', *Behavioral and Brain Sciences*, 3, 111-169.

Pylyshyn, Z.W. (1984). *Computation and cognition. Toward a foundation for cognitive science*, Cambridge Mass., M.I.T. Press.

Quine, W.V.O. (1960). *Word and object*, Cambridge Mass., M.I.T. Press

Rey, G. (1980) 'The formal and the opaque', *Behavioral and Brain Sciences*, 3, 90-92.

Rorty, R. (1970). 'Incorrigibility as the mark of the mental', *Journal of Philosophy*, LXVII (12), 399-424.

Rorty, R. (1971a). 'Mind-body identity, privacy and categories', in Rosenthal, D.M. (ed), *Materialism and the mind-body problem*, Englewood Cliffs N.J., Prentice-Hall.

Rorty, R. (1971b). 'In defense of eliminative materialism', in Rosenthal, D.M. (ed), *Materialism and the mind-body problem*, Englewood Cliffs N.J., Prentice-Hall.

Rorty, R. (1972) 'Functionalism, machines and incorrigibility', *Journal of Philosophy*, LXIX (8), 203-220.

Rorty, R. (1980). *Philosophy and the mirror of nature*, Oxford, Basil Blackwell.

Rorty, R. (1982). 'Contemporary philosophy of mind', *Synthese*, 53, 323-348.

Rosenblith, W.A. (1966). 'On cybernetics and the human brain', *The*

American Scholar, spring 1966.

Rosenthal, D.M. (ed) (1971). *Materialism and the mind-body problem*, Englewood Cliffs, N.J., Prentice-Hall.

Russell, B. (1940). *An inquiry into meaning and truth*, New York, Norton.

Ryle, G. (1949). *The concept of mind*, London, Hutchinson.

Schank, R.C. and Abelson, R.P. (1977). *Scripts, plans, goals and understanding*, Hillsdale, N.J., Erlbaum.

Searle, J. (1969). *Speech acts. An essay in the philosophy of language*, Cambridge, Cambridge University Press.

Searle, J. (1980). 'Minds, brains and programs', *Behavioral and Brain Sciences*, 3, 417-457.

Searle, J. (1983). *Intentionality. An essay in the philosophy of mind*, Cambridge, Cambridge University Press.

Selfridge, O.G. (1955). 'Pattern recognition and modern computers', *Proceedings of the Western Joint Computer Conference*, 7, 91-93.

Sellars, W. (1971). *Science, perception and reality*, London, Routledge & Kegan Paul.

Shannon, C.E. (1956). 'A chess-playing machine', in Newman, J.R. (ed), *The world of mathematics*, New York, Simon & Schuster.

Shoemaker, S. (1975). 'Functionalism and qualia', *Philosophical Studies*, 27, 291-315; herdrukt in Block, N. (ed), *Readings in the philosophy of psychology*, London, Methuen.

Silvers, S. (1985). 'Natural teleology', manuscript, Katholieke Hogeschool Tilburg.

Silvers, S. (1986). Voordracht voor de Katholieke Universiteit Nijmegen, 21 april 1986.

Simon, H.A. and Newell, A. (1958). 'Heuristic problem solving: the next advance in operations research', *Operations Research*, vol. 6, jan.-febr.

Skinner, B.F. (1964). 'Behaviorism at fifty', in Wann, T.W. (ed), *Behaviorism and phenomenology*, Chicago, University of Chicago Press.

Skinner, B.F. (1974). *About behaviorism*, New York, Knopf.

Smart, J.C.C. (1959). 'Sensations and brain processes', *Philosophical Review*, 68, 141-156

Spehlmann, R. (1981). *EEG primer*, Amsterdam, Elsevier/North

Holland

Stabler, E.P. Jr. (1983). 'How are grammars represented?', *Behavioral and Brain Sciences*, 1, 391-423.

Stegmüller, W. (1974). *Probleme und Resultate der Wissenschaftstheorie und analytischen Philosophie. Band I. Wissenschaftliche Erklärung und Begründung*, Berlin, Springer.

Stegmüller, W. (1976). *Hauptströmungen der Gegenwartsphilosophie*, Stuttgart, Alfred Kröner.

Stich, S.P. (1982a). Boekbespreking van Fodor, J.A. *Representations*, in *Contemporary Psychology*, 27 (6), 419-421.

Stich, S.P. (1982b). 'On the ascription of content', in Woodfield, A. (ed), *Thought and object*, Oxford, Clarendon Press.

Stich, S.P. (1983). *From folk psychology to cognitive science*, Cambridge Mass., M.I.T. Press.

Strasser, S. (1965). 'Problemen rondom het begrip der intentionaliteit', *Nederlands Tijdschrift voor de Psychologie en haar grensgebieden*, 20, 1-20.

Straus, E. (1956). *Vom Sinn der Sinne*, Berlin, Springer.

Strawson, P.F. (1959). *Individuals*, London, Methuen & Co.

Strawson, P.F. (1963). 'On referring', in Caton, C.E. (ed), *Philosophy and ordinary language*, Urbana, University of Illinois Press.

Strawson, P.F. (1971). 'Identifying reference and truth values', in Steinberg, D.D. and Jakobovits, L.A. (eds), *Semantics. An interdisciplinary reader in philosophy, linguistics and psychology*, Cambridge, University Press.

Swart, H.A.P. (1985). *Explanation in psychology*, Delft, Eburon.

Tan, Y.-H. (1983). 'Putnam's development from realist to non-realist', Manuscript, Universiteit van Amsterdam.

Thijssen, W.Th.M. (1982). *De mens-machine theorie*, dissertatie Katholieke Universiteit Nijmegen.

Toffler, A. (1971). *Future shock*, New York, Random House

Turing, A.M. (1936). 'On computable numbers, with an application to the Entscheidungsproblem', *Proceedings of the London Mathematical Society*, ser. 2-42, 230-265.

Turing, A.M. (1950) 'Computing machinery and intelligence', *Mind*, LIX (236), 433-460. (Herdrukt in Anderson, A.R. (ed), (1964),

Minds and machines, Englewood Cliffs N.J , Prentice-Hall, 1-30.)

Verwey, G. (1983). 'Kant en Soemmerring. Een confrontatie van wijsbegeerte en neurowetenschap', Voordracht voor de Medisch-Historische Club, KU Nijmegen, 29 november 1983.

Wason, P.C. and Johnson-Laird, P.N. (1972) *Psychology of reasoning: structure and content,* London, Batsford.

Weizenbaum, J. (1965) 'ELIZA - A computer program for the study of natural language communication between man and machine', *Communications of the Association for Computing Machinery,* 9, 1, 36-45.

Weizenbaum, J. (1976). *Computer power and human reason,* San Francisco, Freeman.

Weizenbaum, J. (1984). *Computerkracht en mensenmacht: van oordeel tot berekening,* (vertaling van Computer power and human reason), Amsterdam, Contact.

Winograd, T. (1971). *Procedures as a representation for data in a computer program for understanding natural language,* dissertatie M.I T.

Winograd, T. (1980). 'What does it mean to understand a language', *Cognitive Science,* 4 (3), 209-241.

Wittgenstein, L. (1958). *Philosophical investigations,* Oxford, Basil Blackwell.

Woodfield, A. (ed). (1982). *Thought and object,* Oxford, Clarendon Press.

Woods, W. (1975). 'What's in a link?', in Bobrow, D.G. and Collins, A. (eds), *Representation and understanding,* New York, Academic Press.

Wooldridge, D.E. (1963). *The machinery of the brain,* New York, McGraw-Hill.

M A M M Meijsing, geboren in 1954 deed in 1972 eindexamen gymnasium beta aan de scholengemeenschap Sancta Maria te Haarlem Ze studeerde psychologie en filosofie aan de Universiteit van Amsterdam, behaalde het kandidaatsexamen Psychologie in juni 1975, en het aanvullend kandidaatsexamen Filosofie in juni 1977 Ze was kandidaatsassistent bij de vakgroep Filosofische Anthropologie en Sociale Filosofie van de Universiteit van Amsterdam Na het behalen van het doctoraalexamen Psychologie met als hoofdvak Functieleer in januari 1979 was ze als wetenschappelijk ambtenaar verbonden aan de vakgroep Moderne Letterkunde en aan de vakgroep Wetenschapsfilosofie en Taalfilosofie van de Universiteit van Amsterdam Sedert januari 1981 is ze wetenschappelijk medewerkster bij de vakgroep Wijsgerige Antropologie van de Katholieke Universiteit Nijmegen

STELLINGEN BEHORENDE BIJ HET PROEFSCHRIFT VAN M.A.M.M. MEIJSING.

1. Ofschoon het rationeel kan zijn een gebrek aan empirische progressie in een bepaald researchprogramma te vergoelijken met een beroep op de voortreffelijkheid van de filosofische uitgangspunten, of om een inconsistentie in de filosofische uitgangspunten te vergoelijken met een beroep op de empirische progressie, valt niet in te zien met welk recht het gebruik van beide immuniserende manoeuvres tegelijkertijd (zoals in de cognitiewetenschap wel voorkomt) nog rationeel genoemd kan worden. (Dit proefschrift, hfst. 2).

2. Searle's vroege ('niet-ontologische') werk over intentionaliteit vormt er een - door Searle ongewilde - illustratie van dat de onmogelijkheid een bepaald begrip in fysicalistische termen te analyseren nog niet betekent dat het überhaupt geen wetenschappelijke relevantie heeft. (J.Searle, 'What is an intentional state?' in: H.L. Dreyfus (ed): *Husserl, intentionality and cognitive science,* Cambridge Mass., MIT Press, 1982 en J. Searle, *Intentionality,* Cambridge, University Press, 1983).

3. Pribram's holonomische model van perceptie steunt weliswaar goed zijn anti-localisationistische en constructivistische positie, maar levert (tegengesteld aan wat hij zelf meent) geen model van *perceptie.* (K.H. Pribram, 'Mind, it does matter', in: S.F. Spicker and H.T. Engelhardt (eds): *Philosophical dimensions of the neuro-medical sciences,* Dordrecht, Reidel, 1976).

4. Men kan zich afvragen of Dennett zich met dezelfde gemoedsrust een verificationist zou noemen als hij Popper's argument tegen het verificationisme zou kennen (als we twee bankbiljetten van onachterhaalbare herkomst vinden die, tot en met het serienummer, identiek zijn, weten we zeker dat tenminste één van beide vals is). (K.R. Popper, *Conjectures and refutations,* London, Routledge and

Kegan Paul, 1963)

5. Rorty ondergraaft zijn eigen positie door zijn voorbeeld van de 'Antipodeans', die geen sensaties hebben, maar waarvan er een zegt dat het hebben van vurende C-vezels "just awful" is. (R. Rorty, *Philosophy and the mirror of nature*, Oxford, Basil Blackwell, 1980).

6. De scientistische opstelling die *alle* kritiek op de AI en *alle* bezorgdheid over de invloed van AI op de cultuur afdoet als angst voor (zelf)kennis en voor "a full explanation of man's ability to think". gaat er ten onrechte van uit dat 1) alle kritiek op en bezorgdheid over de AI door angst is ingegeven en dat 2) alle angst per definitie niet rationeel gemotiveerd is. (H. Simon, 'What computers mean for man and society', *Science* 195, march 1977).

7. De systeembenadering in de psychotherapie, volgens welke ieder onderdeel van een systeem er belang bij heeft dat systeem in stand te houden en tot die instandhouding bijdraagt, schiet in gevallen van kindermishandeling en incest fundamenteel tekort.

8. Het valt te vrezen dat het gebruik van uittrekselboeken in het literatuuronderwijs op middelbare scholen, met hun nadruk op het per boek kunnen reproduceren van één eenduidige fabel en één thema, minder het lezen van 'echte' literatuur dan het lezen van triviaalliteratuur zal bevorderen.

9. Zelfs al zou het waar zijn dat Faulkner's romanwereld, zoals de New York Times ooit heeft beweerd, vicious, depraved, decadent, corrupt" is, dan nog volgt daar in het geheel niet uit dat Faulkner zijn oproep, gedaan in zijn dankrede bij het aanvaarden van de Nobelprijs, tot het hooghouden van de oude, universele waarden van "liefde en eer, mededogen en trots, erbarmen en opoffering", ironisch bedoeld heeft. (T. Anbeek, 'Zoals het is, is het al erg genoeg', NRC, CS, 13-12-1985)

10. Het gevaar van positieve discriminatie bij aanstellingen is (*pace*

Edwin Meese, VS minister van Justitie) niet zozeer dat anderen erdoor benadeeld en gestraft worden, maar dat de positief gediscrimineerden zelf onvoldoende kunnen bewijzen dat zij hun aanstelling op grond van eigen kwaliteiten gekregen hebben. ('Hof VS steunt positieve discriminatie', NRC, 3-7-1986).

11 Wanneer men menselijke wezens wil creeren is (over het algemeen) nog steeds de oude beproefde manier bevredigender dan de methoden die sommige AI-wetenschappers voorstaan, zowel voor wat betreft de uitvoering als het resultaat.

SWETS & ZEITLINGER BV